PATENT ABSTRACTS OF JAPAN

(11)Publication number : 10-240747
(43)Date of publication of application : 11.09.1998

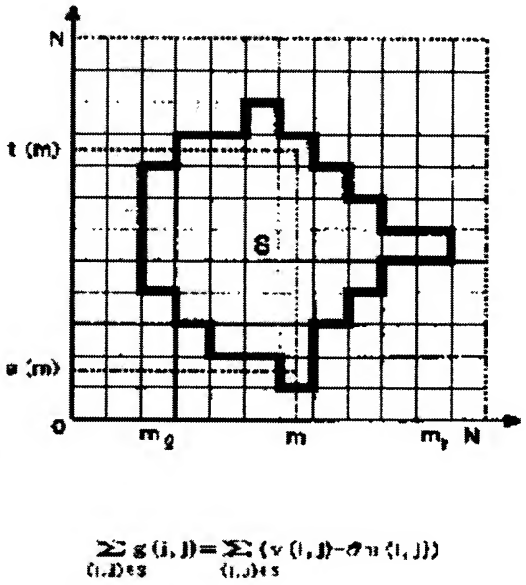
(51)Int.CI. G06F 17/30
G06F 17/18

(21)Application number : 09-034605
(22)Date of filing : 19.02.1997

(71)Applicant : INTERNATL BUSINESS MACH CORP <IBM>
(72)Inventor : YODA KUNIKAZU
FUKUDA TSUYOSHI
TOKUYAMA TAKESHI
MORISHITA SHINICHI

(54) METHOD AND DEVICE FOR DERIVING INTER-DATA CONNECTION RULE, METHOD AND DEVICE FOR CUTTING ORTHOGONAL PROJECTION AREA

(57)Abstract:
PROBLEM TO BE SOLVED: To make it easy to find a connection rule between data by composing a plane of two numeric attributes that an analyzed body has and cutting an orthogonal projection area between true and false attributes under specific conditions.
SOLUTION: A database having data including two kinds of numeric attribute and one kind of true/false attribute has two axes corresponding to two kinds of numeric attribute and stores the number v(i, j) of data belonging to respective pixels (i row, j column) of a plane divided into N × N pixels and the number v(i, j) of data whose true/false attributes are true. Then a specific condition θis inputted to cut an orthogonal projection area S of pixels, maximizing an equation I, out of a plane. Thus, the area in the orthogonal projection shape is cut out to make it easy for people to grasp the connection rule. Lastly, data included in the cut orthogonal projection area S are outputted.



$$\sum_{(i,j)\in S} g(i,j) = \sum_{(i,j)\in S} (v(i,j) - \theta\, u(i,j))$$

LEGAL STATUS

[Date of request for examination]	30.07.1998
[Date of sending the examiner's decision of rejection]	
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]	
[Date of final disposal for application]	
[Patent number]	3193658
[Date of registration]	25.05.2001
[Number of appeal against examiner's decision	

(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号

特開平10－240747

(43)公開日　平成10年(1998)９月11日

(51)Int.Cl.[6]	識別記号	ＦＩ	
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ　15/401	３２０Ｚ
17/18		15/36	Ｚ

審査請求　未請求　請求項の数20　ＯＬ　（全 31 頁）

(21)出願番号　特願平9－34605

(22)出願日　平成９年(1997)２月19日

(71)出願人　390009531
インターナショナル・ビジネス・マシーンズ・コーポレイション
ＩＮＴＥＲＮＡＴＩＯＮＡＬ　ＢＵＳＩＮＥＳＳ　ＭＡＳＣＨＩＮＥＳ　ＣＯＲＰＯＲＡＴＩＯＮ
アメリカ合衆国10504、ニューヨーク州　アーモンク　（番地なし）

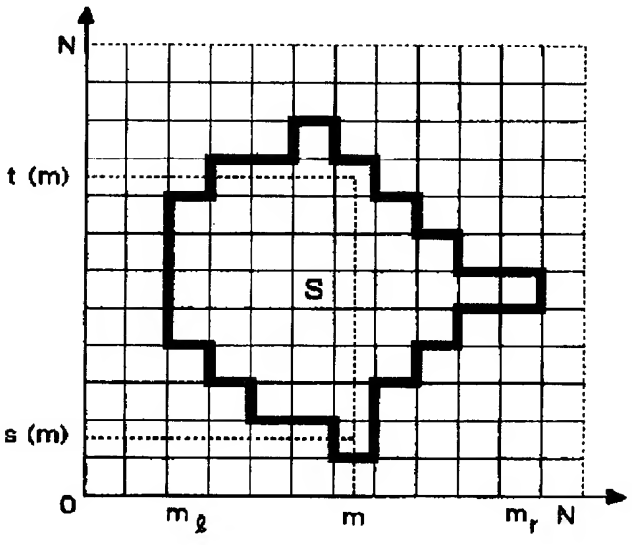
(72)発明者　依田　邦和
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内

(74)代理人　弁理士　合田　潔　（外２名）

最終頁に続く

(54)【発明の名称】　データ間結合ルール導出方法及び装置、及び直交凸領域切出方法及び装置

(57)【要約】
【課題】２項の数値属性と真偽をとる属性を有するデータ間の結合ルールを人間がより把握しやすい形で提示すること。
【解決手段】（１）２つの数値属性により平面を構成し、この平面をピクセルに分割し、各ピクセル内のデータ数及び真偽をとる属性が真となったデータの数をカウントする。（２）所定の条件θに従い、平面の２つの軸に凸な領域である直交凸領域（rectilinear region）を切り出し、データ間の結合ルールを見い出す。（３）切り出した直交凸領域が、先に述べたようなサポート最大化ルール等の条件を満たしていれば、その直交凸領域をユーザに提示する。また、データベースからその直交凸領域に含まれるデータの必要な属性を引き出すことも必要に応じて行う。



1

【特許請求の範囲】

【請求項１】２種類の数値属性と、１種類の真偽をとる属性とを含むデータを有するデータベースにおいて、データ間の結合ルールを導き出す方法であって、

前記２種類の数値属性に対応する２つの軸を有し且つＮ×Ｍ個のピクセルに分割されている平面の各ピクセルに対応して、当該ピクセル（ｉ行ｊ列）に属するデータの数ｕ(ｉ,ｊ)及び前記真偽をとる属性が真であるデータの数ｖ(ｉ,ｊ)を記憶する平面構成ステップと、

条件θを入力するステップと、

【数１】

$$\sum_{(i,j)\in S} g(i,j)=\sum_{(i,j)\in S}(v(i,j)-\theta u(i,j))$$

を最大にするような前記ピクセルの直交凸領域Ｓを前記平面から切り出す領域切出ステップと、

切り出された前記直交凸領域Ｓ内に含まれるデータを出力するステップとを含むデータ間結合ルール導出方法。

【請求項２】入力された前記条件θとは異なる第２の条件θ_2を入力するステップと、

【数２】

$$\sum_{(i,j)\in S_2} g(i,j)=\sum_{(i,j)\in S_2}(v(i,j)-\theta_2 u(i,j))$$

を最大にするような前記ピクセルの第２の直交凸領域S_2を前記平面から切り出すステップと、

【数３】

$$\theta 3=\frac{V(S_2)-V(S)}{U(S_2)-U(S)}$$

（前記直交凸領域S_2に含まれ且つ前記真偽をとる属性が真であるデータの数をＶ（S_2）、前記直交凸領域Ｓに含まれ且つ前記真偽をとる属性が真であるデータの数をＶ（Ｓ）、前記直交凸領域S_2に含まれるデータ数をＵ（S_2）、前記直交凸領域Ｓに含まれるデータ数をＵ（Ｓ）とする。）を第３の条件として、

【数４】

$$\sum_{(i,j)\in S_3} g(i,j)=\sum_{(i,j)\in S_3}(v(i,j)-\theta_3 u(i,j))$$

を最大にするような前記ピクセルの第３の直交凸領域S_3を前記平面から切り出すステップとをさらに含む請求項１記載のデータ間結合ルール導出方法。

【請求項３】前記切り出された直交凸領域Ｓ内の各ピクセルのｖ(ｉ,ｊ)／ｕ(ｉ,ｊ)が、前記平面全体のデータ数に対する前記平面全体の前記真偽をとる属性が真であるデータ数の割合に等しくなるようｖ(ｉ,ｊ)を変更するステップと、

当該変更されたｖ(ｉ,ｊ)を用いて、入力された条件θ_4に従い、

【数５】

2

$$\sum_{(i,j)\in S_4} g(i,j)=\sum_{(i,j)\in S_4}(v(i,j)-\theta_4 u(i,j))$$

を最大にするような前記ピクセルの第４の直交凸領域S_4を切り出すステップとをさらに含む請求項１記載のデータ間結合ルール導出方法。

【請求項４】前記平面構成ステップが、

複数の前記データから、Ｘ個のデータをランダムサンプリングするステップと、

サンプリングされたデータを各前記数値属性についてソートし、Ｘ・ｉ／Ｎ（ｉ＝１，２，・・Ｎ）番目に該当する数値及びＸ・ｎ／Ｍ（ｎ＝１，２，・・Ｍ）番目に該当する数値を記憶するステップと、

記憶された前記数値を基準にして、前記複数のデータの各々がＮ×Ｍ個の前記ピクセルのいずれに含まれるか判断し、各ピクセルにおける数を計数するステップとを含む請求項１記載のデータ間結合ルール導出方法。

【請求項５】前記領域切出ステップが、

第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］が前記区間［ｓ，ｔ］に含まれる直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数１の値が最も大きい直交凸領域$S_m^N(s,t)$の前記第ｍ－１列の区間［ｘ，ｙ］又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^Nに記憶し、当該直交凸領域$S_m^N(s,t)$の数１の値を記憶する第１記憶ステップと、

第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］がｓ≧ｘ及びｔ≧ｙを満たす直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数１の値が最も大きい直交凸領域$S_m^U(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^Uに記憶し、当該直交凸領域$S_m^U(s,t)$の数１の値を記憶する第２記憶ステップと、

第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］がｓ≦ｘ及びｙ≧ｔを満たす直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数１の値が最も大きい直交凸領域$S_m^D(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^Dに記憶し、当該直交凸領域$S_m^D(s,t)$の数１の値を記憶する第３記憶ステップと、

第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ　１列の区間［ｘ，ｙ］がｘ≦ｓ及びｙ≧ｔを満たす直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数１の値が最も大きい直交凸領

域$S_m^N(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^Nに記憶し、当該直交凸領域$S_m^N(s,t)$の数１の値を記憶する第４記憶ステップと、

全てのｍ及び［ｓ，ｔ］について前記第１乃至第４記憶ステップを実行し、最も大きい数１の値を有する直交凸領域Ｓのｍ及び［ｓ，ｔ］と、対応する記憶手段H^N、H^{II}、H^D、又はH^Nの値とを用いて、直交凸領域Ｓを前記平面から切り出すステップとを含む請求項１記載のデータ間結合ルール導出方法。

【請求項６】各々内部に含まれるポイントの数ｕ(ｉ,ｊ)及び所定の条件を満たしたポイントの数ｖ(ｉ,ｊ)を記憶した複数のセルを含む平面から、

【数６】

$$\sum_{(i,j)\in S} g(i,j)=\sum_{(i,j)\in S}(v(i,j)-\theta u(i,j))$$

を最大とする直交凸領域Ｓを切り出す方法であって、

θを入力するステップと、

第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］が前記区間［ｓ，ｔ］に含まれる直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数６の値が最も大きい直交凸領域$S_m^N(s,t)$の前記第ｍ－１列の区間［ｘ，ｙ］又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^Nに記憶し、当該直交凸領域$S_m^N(s,t)$の数６の値を記憶する第１記憶ステップと、

第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］がｓ≧ｘ及びｔ≧ｙを満たす直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数６の値が最も大きい直交凸領域$S_m^{II}(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^{II}に記憶し、当該直交凸領域$S_m^{II}(s,t)$の数６の値を記憶する第２記憶ステップと、

第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］がｓ≦ｘ及びｙ≧ｔを満たす直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数６の値が最も大きい直交凸領域$S_m^D(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^Dに記憶し、当該直交凸領域$S_m^D(s,t)$の数６の値を記憶する第３記憶ステップと、

第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］がｘ≦ｓ及びｙ≧ｔを満たす直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数６の値が最も大きい直交凸領域$S_m^N(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^Nに記憶し、当該直交凸領域$S_m^N(s,t)$の数６の値を記憶する第４記憶ステップと、

全てのｍ及び［ｓ，ｔ］について前記第１乃至第４記憶ステップを実行し、最も大きい数６の値を有する直交凸領域Ｓのｍ及び［ｓ，ｔ］と、対応する記憶手段H^N、H^{II}、H^D、又はH^Nの値とを用いて、直交凸領域Ｓを前記平面から切り出すステップとを含む直交凸領域切出方法。

【請求項７】前記第２記憶ステップが、

第ｍ列の区間［ｓ，ｔ］が右端列であり且つ第ｍ－１列の区間［ｘ，ｙ］がｓ≧ｘ及びｔ＝ｙを満たし且つ第ｍ－２列の区間［ａ，ｂ］と前記第ｍ－１列の区間［ｘ，ｙ］との関係がａ≧ｘ及びｂ≦ｙ又はａ≦ｘ及びｂ≦ｙである直交凸領域、第ｍ列の区間［ｓ，ｔ］が右端列であり且つ第ｍ－１列の区間［ｘ，ｙ］がｓ≧ｘ及びｓ≦ｙ≦ｔ－１を満たし且つ第ｍ－２列の区間［ａ，ｂ］と前記第ｍ－１列の区間［ｘ，ｙ］との関係がａ≧ｘ及びｂ≦ｙ又はａ≦ｘ及びｂ≦ｙである直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数６の値が最も大きい直交凸領域$S_m^{II}(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^{II}に記憶し、当該直交凸領域$S_m^{II}(s,t)$の数６の値を記憶することを特徴とする請求項６記載の直交凸領域切出方法。

【請求項８】前記第３記憶ステップが、

第ｍ列の区間［ｓ，ｔ］が右端列であり且つ第ｍ－１列の区間［ｘ，ｙ］がｓ＝ｘ及びｔ≧ｙを満たし且つ第ｍ－２列の区間［ａ，ｂ］と前記第ｍ－１列の区間［ｘ，ｙ］との関係がａ≧ｘ及びｂ≦ｙ又はａ≧ｘ及びｂ≧ｙである直交凸領域、第ｍ列の区間［ｓ，ｔ］が右端列であり且つ第ｍ－１列の区間［ｘ，ｙ］がｓ＋１≦ｘ≦ｔ及びｙ≧ｔを満たし且つ第ｍ－２列の区間［ａ，ｂ］と前記第ｍ－１列の区間［ｘ，ｙ］との関係がａ≧ｘ及びｂ≦ｙ又はａ≧ｘ及びｂ≧ｙである直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数６の値が最も大きい直交凸領域$S_m^I(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^Dに記憶し、当該直交凸領域$S_m^D(s,t)$の数６の値を記憶す

ることを特徴とする請求項６記載の直交凸領域切出方法。

【請求項９】前記第４記憶ステップが、

第m列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］が前記区間［ｓ，ｔ］と同一である直交凸領域、第m列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］がｘ≦ｓ及びｙ≧ｔ＋１又はｘ≦ｓ－１及びｙ≧ｔを満たす直交凸領域、又は前記第m列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数６の値が最も大きい直交凸領域$S_{st}^{N}(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第m列が左端列であることを示す情報をm及び［ｓ，ｔ］に対応して記憶手段H^{N}に記憶し、当該直交凸領域$S_{st}^{N}(s,t)$の数６の値を記憶することを特徴とする請求項６記載の直交凸領域切出方法。

【請求項１０】前記直交凸領域Ｓを前記平面から切り出すステップが、

全てのm及び［ｓ，ｔ］について前記第１乃至第４記憶ステップを実行するステップと、

計算された数６の値のうち最大の値を有する直交凸領域Ｓのm及び［ｓ，ｔ］と、対応する記憶手段H^{N}、H^{II}、H^{D}、又はH^{N}とを用いて、第ｍ－１列の区間［ｘ，ｙ］及び前記第ｍ－１列と第ｍ－２列との関係を読み出す第１関係読出ステップと、

前記第ｍ－１列と第ｍ－２列との関係を用いて、前記記憶手段H^{N}、H^{II}、H^{D}、及びH^{N}から対応する記憶手段を選択する選択ステップと、

前記第ｍ－１列の区間［ｘ，ｙ］を用いて選択された記憶手段から第ｍ－２列の区間［ａ，ｂ］及び第ｍ－２列と第ｍ－３列との関係を読み出す第２関係読出ステップと、

前記選択ステップと前記第２関係読出ステップとを、前列との関係が前記左端列であることを示す情報となるまで繰り返すステップとを含む請求項６記載の直交凸領域切出方法。

【請求項１１】２種類の数値属性と、１種類の真偽をとる属性を含むデータを有するデータベースにおいて、

データ間の結合ルールを導き出す装置であって、

前記２種類の数値属性に対応する２つの軸を有し且つＮ×Ｍ個のピクセルに分割されている平面の各ピクセルに対応して、当該ピクセル（ｉ行ｊ列）に属するデータの数ｕ(i,j)及び前記真偽をとる属性が真であるデータの数ｖ(i,j)を記憶する平面構成装置と、

条件θを入力する入力デバイスと、

【数７】

$$\sum_{(i,j)\in S} g(i,j)=\sum_{(i,j)\in S}(v(i,j)-\theta u(i,j))$$

を最大にするような前記ピクセルの直交凸領域Ｓを前記平面から切り出す領域切出装置と、

切り出された前記領域Ｓ内に含まれるデータを出力するデバイスとを有するデータ間結合ルール導出装置。

【請求項１２】前記入力デバイスにより、前記条件θとは異なる第２の条件θ_2を入力し、前記領域切出装置により、前記第２の条件θ_2に対応する第２の直交凸領域S_2を前記平面から切り出した場合に、

【数８】

$$\theta 3=\frac{V(S_2)-V(S)}{U(S_2)-U(S)}$$

（前記直交凸領域S_2に含まれ且つ前記真偽をとる属性が真であるデータの数をＶ（S_2）、前記直交凸領域Ｓに含まれ且つ前記真偽をとる属性が真であるデータの数をＶ（Ｓ）、前記直交凸領域S_2に含まれるデータ数をＵ（S_2）、前記直交凸領域Ｓに含まれるデータ数をＵ（Ｓ）とする。）を第３の条件として前記領域切出装置に出力する手段とをさらに有する請求項１１記載のデータ間結合ルール導出装置。

【請求項１３】前記切り出された直交凸領域Ｓ内の各ピクセルのｖ(i,j)／ｕ(i,j)が、前記平面全体のデータ数に対する前記平面全体の前記真偽をとる属性が真であるデータ数の割合に等しくなるようｖ(i,j)を変更する手段と、

当該変更されたｖ(i,j)及び入力された条件θ_4でもって、前記領域切出装置が動作するように命令する手段とを有する請求項１１記載のデータ間結合ルール導出装置。

【請求項１４】前記平面構成装置が、

複数の前記データから、Ｘ個のデータをランダムサンプリングする手段と、

サンプリングされたデータを各前記数値属性についてソートし、Ｘ・ｉ／Ｎ（ｉ＝１，２，・・Ｎ）番目に該当する数値及びＸ・ｎ／Ｍ（ｎ＝１，２，・・Ｍ）番目に該当する数値を記憶する手段と、

記憶された前記数値を基準にして、前記複数のデータの各々がＮ×Ｍ個の前記ピクセルのいずれに含まれるか判断し、各ピクセルにおける数を計数する手段とを含む請求項１１記載のデータ間結合ルール導出装置。

【請求項１５】各々内部に含まれるポイントの数ｕ(i,j)及び所定の条件を満たしたポイントの数ｖ(i,j)を記憶した複数のセルを含む平面から、

【数９】

$$\sum_{(i,j)\in S} g(i,j)=\sum_{(i,j)\in S}(v(i,j)-\theta u(i,j))$$

を最大とする直交凸領域Ｓを切り出す装置であって、

θを入力する手段と、

第m列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］が前記区間［ｓ，ｔ］に含まれる直交凸領域、又は前記第m列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数９の値が最も大きい直交凸

領域$S_{ts}^{N}(s,t)$の前記第m－１列の区間［x，y］又は第m列が左端列であることを示す情報をm及び［s，t］に対応して記憶手段H^{N}に記憶し、当該直交凸領域$S_{ts}^{N}(s,t)$の数９の値を記憶する第１記憶手段と、

第m列の区間［s，t］が右端列であって第m－１列の区間［x，y］がs≧x及びt≧yを満たす直交凸領域、又は前記第m列の区間［s，t］のみで構成される直交凸領域のうち、前記数９の値が最も大きい直交凸領域$S_{ts}^{U}(s,t)$の（a）前記第m－１列の区間［x，y］及び（b）前記第m－１列の区間［x，y］と第m－２列の区間［a，b］との関係、又は第m列が左端列であることを示す情報をm及び［s，t］に対応して記憶手段H^{U}に記憶し、当該直交凸領域$S_{ts}^{U}(s,t)$の数９の値を記憶する第２記憶手段と、

第m列の区間［s，t］が右端列であって第m－１列の区間［x，y］がs≦x及びy≧tを満たす直交凸領域、又は前記第m列の区間［s，t］のみで構成される直交凸領域のうち、前記数９の値が最も大きい直交凸領域$S_{ts}^{D}(s,t)$の（a）前記第m－１列の区間［x，y］及び（b）前記第m－１列の区間［x，y］と第m－２列の区間［a，b］との関係、又は第m列が左端列であることを示す情報をm及び［s，t］に対応して記憶手段H^{D}に記憶し、当該直交凸領域$S_{ts}^{D}(s,t)$の数９の値を記憶する第３記憶手段と、

第m列の区間［s，t］が右端列であって第m－１列の区間［x，y］がx≦s及びy≧tを満たす直交凸領域、又は前記第m列の区間［s，t］のみで構成される直交凸領域のうち、前記数９の値が最も大きい直交凸領域$S_{ts}^{N}(s,t)$の（a）前記第m－１列の区間［x，y］及び（b）前記第m－１列の区間［x，y］と第m－２列の区間［a，b］との関係、又は第m列が左端列であることを示す情報をm及び［s，t］に対応して記憶手段H^{N}に記憶し、当該直交凸領域$S_{ts}^{N}(s,t)$の数９の値を記憶する第４記憶手段と、

全てのm及び［s，t］について前記第１乃至第４記憶手段を動作させ、最も大きい数９の値を有する直交凸領域Sのm及び［s，t］と、対応する記憶手段H^{N}、H^{U}、H^{D}、又はH^{N}の値とを用いて、直交凸領域Sを前記平面から切り出す手段とを有する直交凸領域切出装置。

【請求項１６】２種類の数値属性と、１種類の真偽をとる属性を含むデータを有するデータベースにおいて、コンピュータにデータ間の結合ルールを導き出させるプログラムを記憶した記憶デバイスであって、

前記プログラムは、

前記２種類の数値属性に対応する２つの軸を有し且つN×M個のバケットに分割されている平面の各ピクセルに対応して、当該ピクセル（i行j列）に属するデータの数u(i,j)及び前記真偽をとる属性が真であるデータの数v(i,j)を記憶する平面構成ステップと、

条件θを入力する入力ステップと、

【数１０】

$$\sum_{(i,j)\in S} g(i,j)=\sum_{(i,j)\in S}(v(i,j)-\theta u(i,j))$$

を最大にするような前記ピクセルの直交凸領域Sを前記平面から切り出す領域切出ステップとをコンピュータに実行させる、記憶デバイス。

【請求項１７】前記プログラムが、

前記条件θとは異なる第２の条件θ_2を入力するステップと、

【数１１】

$$\sum_{(i,j)\in S_2} g(i,j)=\sum_{(i,j)\in S_2}(v(i,j)-\theta_2 u(i,j))$$

を最大とするような前記ピクセルの第２の直交凸領域S_2を前記平面から切り出すステップと、

【数１２】

$$\theta 3=\frac{V(S_2)-V(S)}{U(S_2)-U(S)}$$

（前記直交凸領域S_2に含まれ且つ前記真偽をとる属性が真であるデータの数を$V(S_2)$、前記直交凸領域Sに含まれ且つ前記真偽をとる属性が真であるデータの数をV（S）、前記直交凸領域S_2に含まれるデータ数を$U(S_2)$、前記直交凸領域Sに含まれるデータ数をU（S）とする。）を第３の条件として、

【数１３】

$$\sum_{(i,j)\in S_3} g(i,j)=\sum_{(i,j)\in S_3}(v(i,j)-\theta_3 u(i,j))$$

を最大にするような前記ピクセルの第３の直交凸領域S_3を前記平面から切り出すステップとをコンピュータに実行させる、請求項１６記載の記憶デバイス。

【請求項１８】前記プログラムが、

前記切り出された直交凸領域S内の各ピクセルのv(i,j)，u(i,j)が、前記平面全体のデータ数に対する前記平面全体の前記真偽をとる属性のデータ数の割合に等しくなるようv(i,j)を変更するステップと、

当該変更されたv(i,j)及び入力された条件θ_4に従い、

【数１４】

$$\sum_{(i,j)\in S_4} g(i,j)=\sum_{(i,j)\in S_4}(v(i,j)-\theta_4 u(i,j))$$

を最大にするような前記ピクセルの第４の直交凸領域S_4を切り出すステップとをコンピュータに実行させる、請求項１６記載の記憶デバイス。

【請求項１９】前記平面構成ステップが、

複数の前記データから、X個のデータをランダムサンプリングするステップと、

サンプリングされたデータを各前記数値属性についてソートし、X・i／N（i＝１，２，・・N）番目に該当する数値及びX・n／M（n＝１，２，・・M）番目に

該当する数値を記憶するステップと、
記憶された前記数値を基準にして、前記複数のデータの各々がＮ×Ｍ個の前記ピクセルのいずれに含まれるか判断し、各ピクセルにおける数を計数するステップとを含む請求項１６記載の記憶デバイス。
【請求項２０】各々内部に含まれるポイントの数ｕ(i, j)及び所定の条件を満たしたポイントの数ｖ(i, j)を記憶した複数のセルを含む平面から、
【数１５】

$$\sum_{(i,j)\in S} g(i,j) = \sum_{(i,j)\in S} (v(i,j) - \theta u(i,j))$$

を最大とする直交凸領域Ｓを切り出すプログラムを格納した記憶媒体であって、
前記プログラムは、
θを入力するステップと、
第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］が前記区間［ｓ，ｔ］に含まれる直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数１５の値が最も大きい直交凸領域$S_{*}^{N}(s,t)$の前記第ｍ－１列の区間［ｘ，ｙ］又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^{N}に記憶し、当該直交凸領域$S_{*}^{N}(s,t)$の数１５の値を記憶する第１記憶ステップと、
第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］がｓ≦ｘ及びｔ≦ｙを満たす直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数１５の値が最も大きい直交凸領域$S_{*}^{U}(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^{U}に記憶し、当該直交凸領域$S_{*}^{U}(s,t)$の数１５の値を記憶する第２記憶ステップと、
第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］がｓ≦ｘ及びｙ≧ｔを満たす直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数１５の値が最も大きい直交凸領域$S_{*}^{D}(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^{D}に記憶し、当該直交凸領域$S_{*}^{D}(s,t)$の数１５の値を記憶する第３記憶ステップと、
第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］がｘ≦ｓ及びｙ≧ｔを満たす直交凸領域、又は前記第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、前記数１５の値が最も大きい直交凸領域$S_{*}^{N}(s,t)$の（ａ）前記第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）前記第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^{N}に記憶し、当該直交凸領域$S_{*}^{N}(s,t)$の数１５の値を記憶する第４記憶ステップと、
全てのｍ及び［ｓ，ｔ］について前記第１乃至第４記憶ステップを実行し、最も大きい数１５の値を有する直交凸領域Ｓのｍ及び［ｓ，ｔ］と、対応する記憶手段H^{N}、H^{U}、H^{D}、又はH^{N}の値とを用いて、直交凸領域Ｓを前記平面から切り出すステップとをコンピュータに実行させる、記憶媒体。

【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、データベースにおけるデータ相関の解析（データマイニングという。）に関し、より詳しくは２項の数値属性と１項の真偽をとる属性（真偽をとる条件又は０－１属性ともいう。）を有するデータ間の相関を見い出す手法に関する。
【０００２】
【従来の技術】例えば、銀行の顧客を解析対象とし、流動性預金残高がいくらくらいで且つ年齢が何歳ぐらいの人であれば、定期預金残高が２００万円以上になる人が全体の２０％となるか、といった問題を実際に解くことを考える。この流通性預金残高及び年齢は、整数ではあるが連続数値であり、一方定期預金残高２００万円以上というのは、２００万円以上か未満かという分類になるので、真偽をとる属性を有するものである。真偽をとる属性は、例えば「顧客がクレジットカードを有しているか」や「顧客が男性であるか」といった問題と置き換えることも可能である。このような課題を解決することができれば、銀行はどのような人に、例えば新型の金融商品に関するダイレクトメールを送ればよいか簡単に分かるので、効率的な営業活動が行える。
【０００３】従来、先に述べた真偽をとる属性間の相関を表現するルール（結合ルール、association rule）を高速に抽出するような研究は、データマイニングの分野において行われてきた。例えば、R. Agrawal, T. Imielinski, and A. Swami, "Miningassociation rules between sets of items in large databases" In proceedings of the ACM SIGMOD Conference on Management of data, May 1993. や、R. Agrawal and R. Srikant, "Fast algorithms for mining association rules" InProceedings of the 20th VLDB Conference, 1994. 等がある。
【０００４】また、２項の数値データ間のルールを求める従来手法には、以下のようなものがある。
１．強い線形相関を見い出すために、平面上の直線で、点集合を最適近似するものを探す方法。例えば、最小自乗法、再帰中央法等である。これら方法の欠点は、線形相関しか分からず、しかも相関係数の絶対値が０．５以下の場合に線形相関を用いて各データを予測すると精度が低く、現実にはほとんど役にたたない点にある。

２．弱い大域相関を見い出すために、２次元平面上で正方形、長方形、又は円、楕円で面積に対して多くのデータを含むものを見い出す方法。例えば、計算幾何学アルゴリズムを利用するものである。この場合、計算時間が大きくなってしまうという欠点がある。例えば円の場合、Ｏ（M^3）以上の手間が掛かり得る（Ｏ（M^3）は、オーダーM^3の計算手間がかかることを示す。Ｍはデータ数である。）。また、取り出す相関領域としては決まった形をしたものしか扱うことができない。現実には、決まった形で適切にカバーできる場合は少ない。

３．平面を正方メッシュに分割しておき、たくさんのデータを含むピクセルを取り出す方法。しかし、取り出されたピクセルの集合は連結でなく、バラバラなことが多いので、ルールとして見い出すのは困難である。

【０００５】このような手法を用いると、上記の欠点の他に、データ間の多くのルールのうちで、意味のあるものと無意味なものとの区別が難しいという欠点もある。通常、相関に実用上の意味があるかどうかは人間の判断によらないといけないことが多いが、１．や２．では特殊な相関しか取り出せないので意味ある相関を見逃しやすく、３では出力を人間が見てルールを見い出せない。

【０００６】他の方法としては、平面を正方メッシュに分割しておき、これらのピクセルに関して連結且つｘ単調な領域のうち多くのデータを含む領域を切り出す方法がある（Takeshi Fukuda, Yasuhiko Morimoto, Shinichi Morishita and Takeshi Tokuyama,"Data mining using two-dimensional optimized association rules: Scheme, algorithms, and visualization," In Proceedings of the ACM SIGMODConference on Management of Data, pages 13-23, June 1996 を参照のこと）。ｘ単調とは、列方向には凸であるが、行方向では凸ではないものを言う。この方法は、高速で、一定の意味ある相関を取り出すことができるが、縦方向に激しく揺れる入り組んだ領域を切り出すことが多く、人間が見てどこが強い相関の部分であるか把握しにくい。また、ｘ単調ということで、切り出される領域の形状が、正方メッシュのメッシュの仕方（各ピクセルへのデータの配分の仕方）に大きく依存するという欠点もある。

【０００７】

【発明が解決しようとする課題】本発明は、以上のような点に鑑み、２項以上の数値属性と真偽をとる属性を有するデータ間の結合ルールを見い出すための一手法を提供することを目的とする。

【０００８】また、データ間の結合ルールを人間がより把握しやすい形で提示することも目的である。そして、多くの結合ルールを可視化することにより、使用する人間の選択の幅を増大させ、より重要な結合ルールを見いだすこと可能とすることも目的とする。

【０００９】また、（１）真偽をとる属性が真であるデータの割合がある定められた値以上であって、含まれるデータ数が最大となるようなルールであるサポート最大化ルールや、（２）最低限含まれるデータ数が定められた場合、真偽をとる属性が真であるデータの割合が最大となるようなルールであるコンフィデンス最大化ルール、（３）取り出される領域内部と外部との分割を考えた時に、分割前の情報量と比較した分割後の情報量の増分を最大化するルールである最適化エントロピ・ルール、（４）領域内外の分割を考えた時に、内外の「標準化された真偽の割合の平均からのずれ」の二乗和を最大化するルールである最適化インタクラスバリアンス・ルールを満たすような範囲（領域）を導出可能とすることも目的である。

【００１０】さらに、上記のようなデータ間の結合ルールを高速に実行できるような手法を提供することも目的である。

【００１１】

【課題を解決するための手段】通常、解析対象物は多くの数値属性を有する。この中から２つの数値属性を選び、また、１つの真偽をとる属性について、以下のステップを行うことにより、上記の目的を達成するものである。すなわち、

（１）２つの数値属性により平面を構成し、この平面をピクセルに分割し、各ピクセル内のデータ数及び真偽をとる属性が真となったデータの数をカウントする。このような平面は、データ数が濃淡度、真偽をとる属性が真となるデータの数が彩度に該当するような、複数のピクセルを有する平面画像として捉えることもできる。

（２）所定の条件θに従い、平面の２つの軸に凸な領域である直交凸領域（rectilinear region）を切り出し、データ間の結合ルールを見い出す。

（３）切り出した直交凸領域が、先に述べたようなサポート最大化ルール等の条件を満たしていれば、その直交凸領域をユーザに提示する。また、データベースからその直交凸領域に含まれるデータの必要な属性を引き出すことも、必要に応じて行う。

【００１２】なお、切り出された直交凸領域を、そのままユーザに提示したり、複数の直交凸領域を切り出した場合には、それを動画として可視化することにより、所望の結合ルールを見い出し易くすることもできる。

【００１３】また、一旦直交凸領域を切り出した後に、それ以外の結合ルールを見出すべく、切り出された直交凸領域について、彩度を平均化し、再度切り出しステップを実行することも可能である。

【００１４】最初に述べたような例の場合、流動性預金残高の軸と、年齢の軸を設け、その平面を適当なメッシュに分割する。そして、メッシュの各エレメントであるピクセルについて該当する顧客の数と、定期預金残高２００万円以上の顧客の数をカウントする。そして、例えば顧客全体の２０％が入り且つ定期預金残高２００万円以上である顧客の割合が最大となるような直交凸領域で

ある領域の切り出しを行うことにより、コンフィデンス最大化ルールを得ることができる。

【００１５】また、例えば定期預金残高２００万円以上の顧客割合が１０％で最大の顧客数を有する直交凸領域を切り出すことにより、サポート最大化ルールを得ることができる。

【００１６】以上述べた事項をまとめると、２種類の数値属性と、１種類の真偽をとる属性とを含むデータを有するデータベースにおいて、まず、２種類の数値属性に対応する２つの軸を有し且つＮ×Ｍ個のピクセルに分割されている平面の各ピクセルに対応して、当該ピクセル（ｉ行ｊ列）に属するデータの数ｕ(ｉ,ｊ)及び前記真偽をとる属性が真であるデータの数ｖ(ｉ,ｊ)を記憶する（平面構成ステップ）。次に、所定の条件θを入力する。そして、

【数１６】

$$\sum_{(i,j)\in S} g(i,j)=\sum_{(i,j)\in S}(v(i,j)-\theta u(i,j))$$

を最大にするような、ピクセルの直交凸領域Ｓを平面から切り出す（領域切出ステップ）。このように直交凸な形状の領域を切り出すことにより、より人間に結合ルールが把握しやすいようになる。また、先の平面構成ステップからの依存性が小さくできる。最後に、切り出された直交凸領域Ｓ内に含まれるデータを出力する。このように、ルールに合致するデータを得ることができる。

【００１７】また、入力された条件θとは異なる第２の条件θ_2を入力し、

【数１７】

$$\sum_{(i,j)\in S_2} g(i,j)=\sum_{(i,j)\in S_2}(v(i,j)-\theta_2 u(i,j))$$

を最大にするようなピクセルの第２の直交凸領域S_2を平面から切り出し、さらに、

【数１８】

$$\theta 3=\frac{V(S_2)-V(S)}{U(S_2)-U(S)}$$

（直交凸領域S_2に含まれ且つ真偽をとる属性が真であるデータの数をＶ（S_2）、直交凸領域Ｓに含まれ且つ真偽をとる属性が真であるデータの数をＶ（Ｓ）、直交凸領域S_2に含まれるデータ数をＵ（S_2）、直交凸領域Ｓに含まれるデータ数をＵ（Ｓ）とする。）を第３の条件として、

【数１９】

$$\sum_{(i,j)\in S_3} g(i,j)=\sum_{(i,j)\in S_3}(v(i,j)-\theta_3 u(i,j))$$

を最大にするようなピクセルの第３の領域S_3を平面から切り出すようにすることも考えられる。このような処理は、最初の条件θで、初期の目的のルールを導き出せなかった場合に有用である。通常先に示したサポート最大化ルール、コンフィデンス最大化ルール、最適化エントロピ・ルール、最適化インタクラスバリアンス・ルールといったルールを求める際には、条件θを適当に変化させ、上記のような処理を行うことにより求められる。

【００１８】さらに、切り出された直交凸領域Ｓ内の各ピクセルのｖ(ｉ,ｊ)／ｕ(ｉ,ｊ)が、平面全体のデータ数に対する平面全体の真偽をとる属性が真であるデータ数の割合に等しくなるようｖ(ｉ,ｊ)を変更し、当該変更されたｖ(ｉ,ｊ)を用いて、入力された条件θ_4に従い、

【数２０】

$$\sum_{(i,j)\in S_4} g(i,j)=\sum_{(i,j)\in S_4}(v(i,j)-\theta_4 u(i,j))$$

を最大にするようなバケットの第４の領域S_4を切り出すようにすることも考えられる。このようにすると、二次的な相関ルールを導き出すことができる。

【００１９】また、先の平面構成ステップは、複数のデータから、Ｘ個のデータをランダムサンプリングするステップと、サンプリングされたデータを各数値属性についてソートし、Ｘ・ｉ／Ｎ（ｉ＝１，２，・・Ｎ）番目に該当する数値及びＸ・ｎ／Ｍ（ｎ＝１，２，・・Ｍ）番目に該当する数値を記憶し、記憶された数値を基準にして、複数のデータをＮ×Ｍ個のピクセルに入れるようにすることも考えられる。このようにすると、各行各列にデータを高速にまたほぼ均等に割り振ることができる。

【００２０】領域切出ステップは本発明の主要部分である。ここで、第ｍ列の区間［ｓ，ｔ］が右端列であるような直交凸領域は、第ｍ－１列から第ｍ列に移行する際に、第ｍ－１列の区間［ｘ，ｙ］に比して、（１）広がるか、（２）上昇するか、（３）下降するか、（４）狭まるかの４つの類型に分けられる。そして、４つの類型のうち、最も大きい数１６の値を有する直交凸領域が求めたい領域である。よって、それぞれの類型について最大の数１６の値を有する領域を求めるため、以下のようなステップが実行される。

【００２１】第１の類型のため、第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］が区間［ｓ，ｔ］に含まれる直交凸領域、又は第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、数１６の値が最も大きい直交凸領域$S_m^N(s,t)$の第ｍ－１列の区間［ｘ，ｙ］又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応して記憶手段H^Nに記憶し、当該直交凸領域$S_m^N(s,t)$の数１６の値を記憶する。第２の類型のために、第ｍ列の区間［ｓ，ｔ］が右端列であって第ｍ－１列の区間［ｘ，ｙ］がｓ≧ｘ及びｔ≧ｙを満たす直交凸領域、又は第ｍ列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、数１６の値が最も大きい直交凸領域$S_m^H(s,t)$の（ａ）第ｍ－１列の区間［ｘ，ｙ］及び（ｂ）第ｍ－１列の区間［ｘ，ｙ］と第ｍ－２列の区間［ａ，ｂ］との関係、又は第ｍ列が左端列であることを示す情報をｍ及び［ｓ，ｔ］に対応

して記憶手段H^{II}に記憶し、当該直交凸領域$S_m^{II}(s,t)$の数１６の値を記憶する。

【００２２】第３の類型のため、第m列の区間［ｓ，ｔ］が右端列であって第m－１列の区間［ｘ，ｙ］がｓ≦ｘ及びｙ≧ｔを満たす直交凸領域、又は第m列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、数１６の値が最も大きい直交凸領域$S_m^{D}(s,t)$の（ａ）第m－１列の区間［ｘ，ｙ］及び（ｂ）第m－１列の区間［ｘ，ｙ］と第m－２列の区間［ａ，ｂ］との関係、又は第m列が左端列であることを示す情報をm及び［ｓ，ｔ］に対応して記憶手段H^{D}に記憶し、当該直交凸領域$S_m^{D}(s,t)$の数１６の値を記憶する。最後に、第４の類型のため、第m列の区間［ｓ，ｔ］が右端列であって第m－１列の区間［ｘ，ｙ］がｘ≦ｓ及びｙ≧ｔを満たす直交凸領域、又は第m列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、数１６の値が最も大きい直交凸領域$S_m^{N}(s,t)$の（ａ）第m－１列の区間［ｘ，ｙ］及び（ｂ）第m－１列の区間［ｘ，ｙ］と第m－２列の区間［ａ，ｂ］との関係、又は第m列が左端列であることを示す情報をm及び［ｓ，ｔ］に対応して記憶手段H^{N}に記憶し、当該直交凸領域$S_m^{N}(s,t)$の数１６の値を記憶する。

【００２３】各々のステップは、直交凸という性質を考慮して構成されている。そして、全てのm及び［ｓ，ｔ］について前記第１乃至第４記憶ステップを実行し、最も大きい数１６の値を有する直交凸領域Ｓのm及び［ｓ，ｔ］と、対応する記憶手段H^{N}、H^{II}、H^{D}、又はH^{N}の値とを用いて、直交凸領域Ｓを平面から切り出す。

【００２４】なお、第２の類型のための計算は、より詳しく説明すると、第m列の区間［ｓ，ｔ］が右端列であり且つ第m－１列の区間［ｘ，ｙ］がｓ≧ｘ及びｔ＝ｙを満たし且つ第m－２列の区間［ａ，ｂ］と第m－１列の区間［ｘ，ｙ］との関係がａ≧ｘ及びｂ≦ｙ又はａ≦ｘ及びｂ≦ｙである直交凸領域、第m列の区間［ｓ，ｔ］が右端列であり且つ第m－１列の区間［ｘ，ｙ］がｓ≧ｘ及びｓ≦ｙ≦ｔ－１を満たし且つ第m－２列の区間［ａ，ｂ］と第m－１列の区間［ｘ，ｙ］との関係がａ≧ｘ及びｂ≦ｙ又はａ≦ｘ及びｂ≦ｙである直交凸領域、又は第m列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、数１６の値が最も大きい直交凸領域$S_m^{II}(s,t)$の（ａ）第m－１列の区間［ｘ，ｙ］及び（ｂ）第m－１列の区間［ｘ，ｙ］と第m－２列の区間［ａ，ｂ］との関係、又は第m列が左端列であることを示す情報をm及び［ｓ，ｔ］に対応して記憶手段H^{II}に記憶し、当該直交凸領域$S_m^{II}(s,t)$の数１６の値を記憶する、という処理になる。

【００２５】また、第３の類型のための計算は、より詳しく説明すると、第m列の区間［ｓ，ｔ］が右端列であり且つ第m－１列の区間［ｘ，ｙ］がｓ＝ｘ及びｔ≧ｙを満たし且つ第m－２列の区間［ａ，ｂ］と第m－１列の区間［ｘ，ｙ］との関係がａ≧ｘ及びｂ≦ｙ又はａ≧ｘ及びｂ≧ｙである直交凸領域、第m列の区間［ｓ，ｔ］が右端列であり且つ第m－１列の区間［ｘ，ｙ］がｓ＋１≦ｘ≦ｔ及びｙ≧ｔを満たし且つ第m－２列の区間［ａ，ｂ］と第m－１列の区間［ｘ，ｙ］との関係がａ≧ｘ及びｂ≦ｙ又はａ≧ｘ及びｂ≧ｙである直交凸領域、又は第m列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、数１６の値が最も大きい直交凸領域$S_m^{D}(s,t)$の（ａ）第m－１列の区間［ｘ，ｙ］及び（ｂ）第m－１列の区間［ｘ，ｙ］と第m－２列の区間［ａ，ｂ］との関係、又は第m列が左端列であることを示す情報をm及び［ｓ，ｔ］に対応して記憶手段H^{D}に記憶し、当該直交凸領域$S_m^{D}(s,t)$の数１６の値を記憶する、という処理になる。

【００２６】さらに、第４の類型のための計算は、より詳しく説明すると、第m列の区間［ｓ，ｔ］が右端列であって第m－１列の区間［ｘ，ｙ］が区間［ｓ，ｔ］と同一である直交凸領域、第m列の区間［ｓ，ｔ］が右端列であって第m－１列の区間［ｘ，ｙ］がｘ≦ｓ及びｙ≧ｔ＋１又はｘ≦ｓ－１及びｙ≧ｔを満たす直交凸領域、又は第m列の区間［ｓ，ｔ］のみで構成される直交凸領域のうち、数１６の値が最も大きい直交凸領域$S_m^{N}(s,t)$の（ａ）第m－１列の区間［ｘ，ｙ］及び（ｂ）第m－１列の区間［ｘ，ｙ］と第m－２列の区間［ａ，ｂ］との関係、又は第m列が左端列であることを示す情報をm及び［ｓ，ｔ］に対応して記憶手段H^{N}に記憶し、当該直交凸領域$S_m^{N}(s,t)$の数１６の値を記憶する、という処理になる。

【００２７】また、最後に切り出すステップは、最初に、全てのm及び［ｓ，ｔ］について第１乃至第４記憶ステップを実行し、計算された数１６の値のうち最大の値を有する直交凸領域Ｓのm及び［ｓ，ｔ］と、対応する記憶手段H^{N}、H^{II}、H^{D}、又はH^{N}とを用いて、第m－１列の区間［ｘ，ｙ］及び第m－１列と第m－２列との関係を読み出す（第１関係読出ステップ）。そして、第m－１列と第m－２列との関係を用いて、記憶手段H^{N}、H^{II}、H^{D}、及びH^{N}から対応する記憶手段を選択し（選択ステップ）、第m－１列の区間［ｘ，ｙ］を用いて選択された記憶手段から第m－２列の区間［ａ，ｂ］及び第m－２列と第m－３列との関係を読み出す（第２関係読出ステップ）。最後に選択ステップと第２関係読出ステップとを、前列との関係が「左端列であることを示す情報」となるまで繰り返す。

【００２８】以下の説明を理解すれば、上述の方法を実施するような装置を作成すること、またこのような方法をコンピュータに実施させるプログラムを作成することは容易に実施できるであろう。また、上記のようなプログラムを記憶媒体や記憶デバイスに記憶することは、通常行われることである。

【００２９】
【発明の実施の形態】まず、本発明の各ステップがどのように実施されるかを示す。
（１）平面構成ステップ
先に述べたように、あるデータの２つの数値属性に２つの座標軸（ｘ軸，ｙ軸）をそれぞれ対応させ、これら２軸の張る平面を考える。この平面を軸ごとにＮ個のピクセルに分割し、平面上にＮ^{2}個のピクセルを作成する。図１に、この平面構成ステップのフローを示す。ステップ１００にて処理が開始し、まずデータ集合Ｐからデータのランダムサンプリングを行う（ステップ１１０）。サンプリングされたデータをｐｋ（ｘｋ，ｙｋ）（ｋ＝１，２，．．．Ｘ。ｘｋ，ｙｋはデータの２つの数値属性の値を、Ｘはサンプリングされたデータ数をそれぞれ示す。）と表す。そして、ｘｋ，ｙｋごとにソートを行い（ステップ１２０）、ｘｋ，ｙｋごとに、ｉ・Ｘ／Ｎ（ｉ＝１，２，．．．Ｎ－１）番目に小さな値を見つけ出す（ステップ１３０）。見つけ出された値が、各軸のピクセルの境界値となる。このようなステップを実施することにより、平面上の各列及び各行に属するデータの数はほぼ均等になる。そして、見つけ出された値を用いて、ピクセル（ｉ，ｊ）に入るデータｐｋの数ｕ（ｉ，ｊ）と、その中で真偽をとる属性が真であるデータｐｋの数ｖ（ｉ，ｊ）とをカウントする（ステップ１４０）。ここで、ｕ（ｉ，ｊ）及びｖ（ｉ，ｊ）は、上記平面上のｉ行ｊ列に存在するピクセルのデータを表すので、ｘ軸方向にｊ、ｙ軸方向にｉ進んだ場所にあるピクセルを表すことになる点に注意する。最後に、カウントされたｕ（ｉ，ｊ）及びｖ（ｉ，ｊ）を各ピクセルごとに記憶する（ステップ１５０）。このようにして、２つのＮ×Ｎ行列、ｕ（ｉ，ｊ）及びｖ（ｉ，ｊ）が生成される。
【００３０】上述のようにランダムサンプリングを行うのは、通常全てのデータをソートしていると時間がかかるからである。但し、ソートしてもよいような場合もある。また、ランダムサンプリングで取り出されるデータ＊の数は、３０Ｎから５０Ｎぐらいが好ましい。また、２軸ともＮ個に分割する例を示したが、異なる数に分割することも可能である。典型的な例で、Ｎは２０から１０００ぐらいである。
【００３１】以上述べたのは一例であって、他の方法を用いてもよい。例えば、各ピクセルの境界数値については予め定めた値を用いても良い。また、データ値に対して均等に分割することも、また対数的に分割することも可能である。
【００３２】また、後の処理のため以下のような処理（図２）を行っておくと、さらに全体の処理が高速化される。すなわち、ｕ(ｉ,ｊ)とｖ(ｉ,ｊ)の行数（N_y）と列数（N_x）を調べる（ステップ２１０）。そして、先に求めたｕ(ｉ,ｊ)とｖ(ｉ,ｊ)を用いて、新たに以下のようなｕ'(ｉ,ｊ)とｖ'(ｉ,ｊ)という行列を作成する（ステップ２２０）。
【数２１】

$$u'(i,j)=\sum_{k=0}^{i}u(k,j)$$

【数２２】

$$v'(i,j)=\sum_{k=0}^{i}v(k,j)$$

これらの計算は、全てのｉ＝０，１，．．N_y－１，ｊ＝０，１，．．．N_x－１について実施する。
【００３３】このｕ'(ｉ,ｊ)とｖ'(ｉ,ｊ)は、後々数多く計算することとなる目的関数（以下、ゲインということもある）、
【数２３】

$$g(i,j)=v(i,j)-\theta u(i,j)$$

の和計算を以下のように簡単化するために用意する。
【数２４】

$$\Gamma_m(s,t)=\sum_{k=s}^{t}g(k,m)$$

$$=\begin{cases}v'(t,m)-\theta u'(t,m) & (s=0)\\ (v'(t,m)-v'(s-1,m))-\theta(u'(t,m)-u'(s-1,m)) & (s\geq 1)\end{cases}$$

【００３４】さらに、
【数２５】

$$Usum=\sum_{全体}u$$

【数２６】

$$Vsum=\sum_{全体}v$$

も後によく用いるので用意する。以上のような準備をすれば、以下の領域切り出しステップが高速になる。
【００３５】（２）領域切り出しステップ
このステップは直交凸領域を先に作成した平面から切り出すものである。直交凸領域の例を図３に示す。直交凸領域は、（１）ｙ軸に平行な線との交わりが必ず連続か空であって、且つ（２）ｘ軸に平行な線との交わりが必ず連続か空な領域を言う。図３の左側の領域は、ｙ軸に平行な、いかなる線との交わりも連続か空であり、且つｘ軸に平行な、いかなる線との交わりも連続か空であるので直交凸領域であると言える。一方、図３の右側の領

域は、ｘ軸に平行な線との交わりは必ず連続又は空であるが、ｙ軸に平行な線でとの交わりは連続でないものを含んでいる。よって、この領域は直交凸ではない。

【００３６】先に示した直交凸領域の条件（１）のみを満たす領域をｘ単調な領域と言い、（２）のみを満たす領域はｙ単調な領域という。先に示した従来技術では、ｘ単調な領域を切り出すものであったが、実際に本発明のようなデータマイニングにおいて当該従来技術を適用すると、縦に激しく揺れる入り組んだ領域を切り出すことが多く、人間には理解し難い形状となってしまう。また、任意の形状で切り出そうとすると、その問題はＮＰ困難となってしまう。そこで、本発明では、直交凸領域で切り出すこととする。

【００３７】直交凸領域を切り出す際には、パラメータ値θ（０以上１以下の実数）を含む数２３で表されるゲインを領域全体で最大にするような直交凸領域を切り出す。ここで、パラメータθの説明をしておく。図４に示すような、横軸が切り出される領域Ｓに含まれるデータ数Ｕ（Ｓ）、縦軸が切り出される領域Ｓに含まれ且つ真偽をとる属性が真であるデータの数Ｖ（Ｓ）であるような平面を考える。データ数と真偽をとる属性が真であるデータの数の組み合わせは多数存在するので、この平面には多数の点が存在することになるが、この点のうち、凸包を構成する点を特に用いる。すなわち、この凸包を構成する点をつなぐことにより曲線を構成し、この曲線に対し傾きθを有する直線を上から下ろして行き、最初にこの曲線と接する点を求め、この時の領域を出力する。凸包上の点は図４では黒丸で表されている。以下、凸包上の点をフォーカス・イメージという。また、直線を下ろしていくような方法をハンドプローブという。このように、本発明では傾きθをパラメータとして入力するような方法を用いる。

【００３８】このように凸包上の点のみ取り扱うのは、コンフィデンス最大化ルール、サポート最大化ルールは、凸包上に必ず存在するわけではないが、近似解としては十分な点を出力することができ、また最適化エントロピ・ルール及び最適化インタクラスバリアンス・ルールについては、この凸包上に必ず存在するからである。もし、コンフィデンス最大化ルール及びサポート最大化ルールの厳密解を解くとすると、実用的な時間には計算が終了しないので、近似解であっても十分に有効な結果を出力できる。

【００３９】上記のように傾きθの直線を下ろしていくということは、直線ｙ＝θｘ＋ＱのＹ切片であるＱを減少させることであり、言いかえれば、Ｑ＝Ｖ（Ｓ）－θＵ（Ｓ）を最大にするＵ（Ｓ）をＸ座標に有する点を求める問題となる。よって、

【数２７】

$$\max Q=\max\{\sum_{(i,j)\in S} v(i,j)-\theta\sum_{(i,j)\in S} u(i,j)\}$$

$$=\max\sum_{(i,j)\in S} g(i,j)$$

と変形される。

【００４０】では、この数２７をどのように解くかを考える。最初に、直交凸領域の性質を領域内部のピクセル同士の関係によって表現する。Ｓをピクセル平面内の直交凸領域とする。m_l、m_rをそれぞれＳの左端、右端の列番号とする。Ｓの第ｍ列（$m_l \leqq m \leqq m_r$）の下端及び上端のピクセル番号をそれぞれｓ（ｍ），ｔ（ｍ）とする。これらの位置関係は図５を参照するとよく分かる。第ｍ列の区間［ｓ（ｍ），ｔ（ｍ）］の変化傾向を第ｍ－１列の区間との比較によって次のように定義する。

（ａ）Ｗ－Ｔｙｐｅ：広がり型　（図６左上）
ｓ（ｍ－１）≧ｓ（ｍ），ｔ（ｍ－１）≦ｔ（ｍ）の場合

（ｂ）Ｕ－Ｔｙｐｅ：上昇型　（図６右上）
ｓ（ｍ－１）≦ｓ（ｍ），ｔ（ｍ－１）≦ｔ（ｍ）の場合

（ｃ）Ｄ－Ｔｙｐｅ：下降型　（図６左下）
ｓ（ｍ－１）≧ｓ（ｍ），ｔ（ｍ－１）≧ｔ（ｍ）の場合

（ｄ）Ｎ－Ｔｙｐｅ：狭まり型　（図６右下）
ｓ（ｍ－１）≦ｓ（ｍ），ｔ（ｍ－１）≧ｔ（ｍ）の場合

【００４１】ｍ＝m_lの列は全変化傾向に属し、上の不等式の等号が成り立つ場合、その列は複数の変化傾向に同時に属する。上の定義から直交凸領域内のどの列の区間も上の４種類のタイプのいずれかに属する。

【００４２】また、直交凸領域の性質から、ある変化傾向の列の左隣の列の変化傾向は次の条件を満たす。すなわち、

（１）Ｗ－Ｔｙｐｅの左隣の列はＷ－Ｔｙｐｅである。
（２）Ｕ－Ｔｙｐｅの左隣の列はＷ－Ｔｙｐｅ又はＵ－Ｔｙｐｅである。
（３）Ｄ－Ｔｙｐｅの左隣の列はＷ－Ｔｙｐｅ又はＤ－Ｔｙｐｅである。
（４）Ｎ－Ｔｙｐｅの左隣の列はＷ－Ｔｙｐｅ、又はＵ－Ｔｙｐｅ、又はＤ－Ｔｙｐｅ、又はＮ－Ｔｙｐｅである。

【００４３】このような条件を満たす領域は逆に言うと直交凸領域であると言える。これらの条件は、図７に状態遷移図として示されている。図中のＷ，Ｕ，Ｄ，Ｎは、それぞれＷ－Ｔｙｐｅ、Ｕ－Ｔｙｐｅ、Ｄ－Ｔｙｐｅ、Ｎ－Ｔｙｐｅであり、矢印を１つたどるごとに１つ右隣の列の状態に遷移する。

【００４４】全ての直交凸領域は領域の右端の列の変化傾向によって先に示した４つの種類に分類できる。ここ

21

で、４つの種類を総称してＸタイプ（Ｘ∈ {Ｗ，Ｕ，Ｄ，Ｎ}）と呼ぶ。列の区間のタイプと同様に、領域のタイプも複数のタイプに同時に属する場合もある。

【００４５】また、右端が第ｍ列の区間［ｓ，ｔ］であ＊

22

＊るＸタイプの直交凸領域のゲインの最大値を$f_m^X(s,t)$と表す。そして、４つのタイプ領域のゲインのうち最も大きいものを、$f_m(s,t)$と表す。すなわち、

【数２８】

$$f_m(s,t) = \max\{f_m^W(s,t), f_m^U(s,t), f_m^D(s,t), f_m^N(s,t)\}$$

である。この$f_m(s,t)$をｍ＝０，．．．N_x-1（∀（ｓ≦ｔ））について求めて、それらの中で最大のものを選びだせば、それが先の平面内の全ての直交凸領域のゲインの最大値となる。

【００４６】この最大値を求めるために、ｍ＝０．．．N_x-1に対して順番に、$f_m(s,t)$（∀（ｓ≦ｔ））を全て計算するという方針をとる。

【００４７】次にｍ＝０、すなわち第１列の$f_0^X(s,t)$を計算する。この場合、全てのタイプで同一である。これは、

【数２９】

$$f_0^X(s,t)=\Gamma_0(s,t) \quad \forall X\in\{W,U,D,N\} \quad \forall(s\leqq t)$$

※で求められる。

【００４８】そして、$f_{m-1}^X(s,t)$（∀Ｘ∈ {Ｗ，Ｕ，Ｄ，Ｎ}，（∀（ｓ≦ｔ））を求める。以下は、各タイプごとに説明する。

【００４９】（ａ）広がり型（Ｗ－Ｔｙｐｅ）の場合

第ｍ列の区間［ｓ，ｔ］を最右端とする直交凸領域であって第ｍ列がＷ－Ｔｙｐｅである領域のゲインの最大値$f_m^W(s,t)$は、以下の式により求められる。

【数３０】

※

$$f_m^W(s,t)=\max\begin{cases}\Gamma_m(s,t) & (1)\\ f_{m-1}^W(s,t)+\Gamma_m(s,t) & (2)\\ f_m^W(s,t-1)+g(t,m) \quad (s<t) & (3)\\ f_m^W(s+1,t)+g(s,m) \quad (s<t) & (4)\end{cases}$$

ここで、ｍａｘを求める時に、ｓ＝ｔの場合は数３０の（１）（２）式だけで比較をし、大きい方の値を用いる。その他の場合は（１）乃至（４）のすべてから最も大きい値を用いる。

【００５０】数３０の（１）式は、第ｍ列の区間［ｓ，ｔ］だけからなる領域（幅１の縦長の長方形）のゲインを表す。また（２）式は、第ｍ－１列がＷ－Ｔｙｐｅでその区間が［ｓ，ｔ］、且つ第ｍ列も区間［ｓ，ｔ］で右端となっている領域のうち最大のゲインを表す。これは、図８（ａ）に表したような場合を示す。なお、直交凸領域の性質から第ｍ列がＷ－Ｔｙｐｅであれば第ｍ－１列がＷ－Ｔｙｐｅであることは決まる。

【００５１】また（３）式は、第ｍ－１列がＷ－Ｔｙｐｅであって、その区間［ｓ（ｍ－１），ｔ（ｍ－１）］が、ｓ（ｍ－１）≧ｓ，ｔ（ｍ－１）≦ｔ－１を満たし、第ｍ列の区間［ｓ，ｔ］で右端という領域のうち最大のゲインを示す。これは、図８（ｂ）のような形状を意味する。第ｍ－１列の上端は、ｔ－１以下であり、下端はｓ以上である。（４）式は、第ｍ－１列がＷ－Ｔｙｐｅであって、その区間［ｓ（ｍ－１），ｔ（ｍ－１）］がｓ（ｍ－１）≧ｓ＋１，ｔ（ｍ－１）≦ｔを満たし、第ｍ列の区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは、図８（ｃ）のような形状を意味する。第ｍ－１列の上端はｔ以下であり、下端はｓ＋１以上である。（２）乃至（４）式は、Ｗ－Ｔｙｐｅの左列はＷ－Ｔｙｐｅしかあり得ないということが考慮されている。

【００５２】以上の$f_m^W(s,t)$の計算を一列中の全ての区間［ｓ，ｔ］に対して行う。この計算は図９のアルゴリズムに従う。以上のように、第ｍ列が領域の右端で第ｍ－１列からの変化傾向がＷ－Ｔｙｐｅという領域のうち最大のゲインが得られる。

【００５３】（ｂ）上昇型（Ｕ－Ｔｙｐｅ）の場合

最初に以下の式の値を求めておく。

【数３１】

$$\beta_{m-1}^W(s,t)=\{i \mid \max_{i\leqq s} f_{m-1}^W(i,t)\}$$

【数３２】

23

$$\beta^{U}_{m-1}(s,t)=\{i \mid \max_{i\leq s} f^{U}_{m-1}(i,t)\}$$

これは、図１０のようなアルゴリズムにて実行される。

【００５４】以上の計算を用いて、第m列の区間［ｓ，＊

24

＊ｔ］を最右端とする直交凸領域であって第m列がＵ－Ｔｙｐｅである領域のゲインの最大値$f_m^{"}(s,t)$は、以下の式により求められる。

【数３３】

$$f^{U}_{m}(s,t)=\max\begin{cases}\Gamma_m(s,t) & (1)\\ f^{W}_{m-1}(\beta^{W}_{m-1}(s,t),t)+\Gamma_m(s,t) & (2)\\ f^{U}_{m-1}(\beta^{U}_{m-1}(s,t),t)+\Gamma_m(s,t) & (3)\\ f^{U}_{m}(s,t-1)+g(t,m) \quad (s<t) & (4)\end{cases}$$

（１）乃至（３）式は、ｓ＝ｔの場合に比較に用いられ、その際（４）式は用いられない。

【００５５】数３３の（１）式は、第m列の区間［ｓ，ｔ］のみからなる領域（幅１の縦長の長方形）のゲインを表す。また、（２）式は、第m－１列がＷ－Ｔｙｐｅであって、その区間［ｓ（ｍ－１），ｔ（ｍ－１）］が、ｓ（ｍ－１）≦ｓ，ｔ（ｍ－１）＝ｔを満たし、第m列は区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは、図１１（ａ）に示した形状の場合であって、第m－１列の下端の上限はｓである。

【００５６】（３）式は、第m－１列がＵ－Ｔｙｐｅであって、その区間［ｓ（ｍ－１），ｔ（ｍ－１）］が、ｓ（ｍ－１）≦ｓ，ｔ（ｍ－１）＝ｔを満たし、第m列は区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは、図１１（ｂ）に示した形状の場合であって、第m－１列の下端の上限はｓである。（４）式は、第m－１列がＷ－Ｔｙｐｅ又はＵ－Ｔｙｐｅであって、その区間［ｓ（ｍ－１），ｔ（ｍ－１）］が、ｓ（ｍ－１）≦ｓ，ｓ≦ｔ（ｍ－１）≦ｔ－１を満たし、第m列は区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは、図１１（ｃ）に示した形状の場合であって、第m－１列の下端の上限はｓであり、また上端の範囲はｓ以上ｔ－１以下である。（２）乃至（４）式は、Ｕ－Ｔｙｐｅの左隣列はＷ－Ｔｙｐｅ又はＵ－Ｔｙｐｅしかあり得ないということが考慮されてい※る。

【００５７】以上の$f_m^{"}(s,t)$の計算を一列中の全ての区間［ｓ，ｔ］に対して行う。この計算は、図１２に示すアルゴリズムに従う。このように、第m列が領域の右端でそこの変化傾向がＵ－Ｔｙｐｅという領域のうち最大のゲインが得られる。

【００５８】（ｃ）下降型（Ｄ－Ｔｙｐｅ）の場合

最初に以下の式の値を計算しておく。

【数３４】

$$\tau^{W}_{m-1}(s,t)=\{i \mid \max_{i\geq t} f^{W}_{m-1}(s,i)\}$$

【数３５】

$$\tau^{D}_{m-1}(s,t)=\{i \mid \max_{i\geq t} f^{D}_{m-1}(s,i)\}$$

これらの計算は、図１３に示されたアルゴリズムにより実行される。（１）乃至（３）式は、ｓ＝ｔの場合に比較に用いられ、その際（４）式は用いられない。

【００５９】以上の計算を用いて、第m列の区間［ｓ，ｔ］を最右端とする直交凸領域であって第m列がＤ－Ｔｙｐｅである領域のゲインの最大値$f_m^{'}(s,t)$は、以下の式により求められる。

【数３６】

$$f^{D}_{m}(s,t)=\max\begin{cases}\Gamma_m(s,t) & (1)\\ f^{W}_{m-1}(s,\tau^{W}_{m-1}(s,t))+\Gamma_m(s,t) & (2)\\ f^{D}_{m-1}(s,\tau^{D}_{m-1}(s,t))+\Gamma_m(s,t) & (3)\\ f^{D}_{m}(s+1,t)+g(s,m) \quad (s<t) & (4)\end{cases}$$

【００６０】数３６の（１）式は、第m列の区間［ｓ，ｔ］のみからなる領域（幅１の縦長の長方形）のゲインを表す。また、（２）式は、第m－１列がＷ－Ｔｙｐｅであって、その区間［ｓ（ｍ－１），ｔ（ｍ－１）］が、ｓ（ｍ－１）＝ｓ，ｔ（ｍ－１）≧ｔを満たし、第m列は区間［ｓ，ｔ］で右端という領域のうち最大のゲ

25

インを表す。これは、図１４（ａ）に示した形状の場合であって、第ｍ－１列の上端の下限はｔである。

【００６１】（３）式は、第ｍ－１列がＤ－Ｔｙｐｅであって、その区間［ｓ（ｍ－１），ｔ（ｍ－１）］が、ｓ（ｍ－１）＝ｓ，ｔ（ｍ－１）≧ｔを満たし、第ｍ列は区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは図１４（ｂ）に示した形状の場合であって、第ｍ－１列の上端の下限はｔである。（４）式は、第ｍ－１列がＷ－Ｔｙｐｅ又はＤ－Ｔｙｐｅであって、その区間［ｓ（ｍ－１），ｔ（ｍ－１）］が、ｓ＋１≦ｓ（ｍ－１）≦ｔ，ｔ（ｍ－１）≧ｔを満たし、第ｍ列は区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは図１４（ｃ）に示した形状の場合であって、第ｍ－１列の上端の下限はｔであって、下端の範囲＊

26

＊はｓ＋１以上ｔ以下である。（２）乃至（４）式は、Ｄ－Ｔｙｐｅの左隣列はＷ－Ｔｙｐｅ又はＤ－Ｔｙｐｅしかあり得ないということが考慮されている。

【００６２】以上の$f_m^P(s,t)$の計算を一列中の全ての区間［ｓ，ｔ］に対して行う。この計算は図１５のアルゴリズムに従う。このようにして、第ｍ列が領域の右端でそこの変化傾向がＤ－Ｔｙｐｅという領域のうち最大のゲインが得られる。

【００６３】（ｄ）狭まり型（Ｎ－Ｔｙｐｅ）の場合第ｍ列の区間［ｓ，ｔ］を最右端とする直交凸領域であって第ｍ列がＮ－Ｔｙｐｅである領域のゲインの最大値$f_m^N(s,t)$は、以下の式により求められる。

【数３７】

$$f_m^N(s,t)=\max\begin{cases}\Gamma_m(s,t) & (1)\\ f_{m-1}^W(s,t)+\Gamma_m(s,t) & (2)\\ f_{m-1}^U(s,t)+\Gamma_m(s,t) & (3)\\ f_{m-1}^D(s,t)+\Gamma_m(s,t) & (4)\\ f_m^N(s,t+1)+\Gamma_m(s,t) & (5)\\ f_m^N(s,t+1)-g(t+1,m) \quad (t<N_y-1) & (6)\\ f_m^N(s-1,t)-g(s-1,m) \quad (s>0) & (7)\end{cases}$$

ここで、ｍａｘを求める時、各式は式の後ろの条件を満たす場合にのみ用いられる。すなわち、（６）式はｔ＞N_y-1を満たす時のみ比較され、（７）式はｓ＞０を満たす場合にのみ比較に用いられる。

【００６４】数３７の（１）式は、第ｍ列の区間［ｓ，ｔ］のみからなる領域（幅１の縦長の長方形）のゲインを表す。（２）式は、第ｍ－１列がＷ－Ｔｙｐｅであって、その区間が［ｓ，ｔ］であり、第ｍ列は区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは図１６（ａ）に示した形状の場合である。（３）式は、第ｍ－１列の区間［ｓ，ｔ］がＵ－Ｔｙｐｅであって、第ｍ列は区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは図１６（ｂ）に示した形状の場合である。（４）式は、第ｍ－１列の区間［ｓ，ｔ］がＤ－Ｔｙｐｅであって、第ｍ列は区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは図１６（ｃ）に示した形状の場合である。（５）式は、第ｍ－１列の区間［ｓ，ｔ］がＮ－Ｔｙｐｅであって、その区間［ｓ，ｔ］であり、第ｍ列は区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは図１６（ｄ）に示した形状の場合である。

【００６５】（６）式は、第ｍ－１列がＷ－Ｔｙｐｅ、Ｕ－Ｔｙｐｅ、Ｄ　Ｔｙｐｅ又はＮ－Ｔｙｐｅであって、その区間［ｓ（ｍ－１），ｔ（ｍ－１）］が、ｓ（ｍ－１）≦ｓ，ｔ（ｍ－１）≧ｔ＋１を満たし、第ｍ列は区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは図１６（ｅ）に示した形状であって、第ｍ－１列の上端の下限はｔ＋１であり、下端の上限はｓである。（７）式は、第ｍ－１列がＷ－Ｔｙｐｅ、Ｕ－Ｔｙｐｅ、Ｄ－Ｔｙｐｅ又はＮ－Ｔｙｐｅであって、その区間［ｓ（ｍ－１），ｔ（ｍ－１）］が、ｓ（ｍ－１）≦ｓ－１，ｔ（ｍ－１）≧ｔを満たし、第ｍ列は区間［ｓ，ｔ］で右端という領域のうち最大のゲインを表す。これは図１６（ｆ）に示した形状であって、第ｍ－１列の上端の下限はｔであり、下端の上限はｓ－１である。

【００６６】以上の$f_m^N(s,t)$の計算を一列中の全ての区間［ｓ，ｔ］に対して行う。この計算は次のアルゴリズムに図１７に従う。このようにして、第ｍ列の区間［ｓ，ｔ］が領域の右端でそこの変化傾向がＮ－Ｔｙｐｅという領域のうち最大のゲインが得られる。

【００６７】上述の（ａ）乃至（ｄ）の計算にて各列の

各［ｓ，ｔ］を右端とする領域の最大のゲインを計算することができる訳であるが、それと同時に"領域"自体も同時に記録しておく必要がある。これは、後の出力ステップでは、この求められた直交凸領域内に含まれるデータを取り出すからである。

【００６８】ここで、同じ最大値の領域が複数存在する場合には、それらのうち先に見つかった方を解として取り扱う。また、領域は縦方向の区間が横に並んだものとして

［ｓ（m_1），ｔ（m_1）］，．．．［ｓ（m_r），ｔ（m_r）］

のように表現する。１列はN_y行あるので、区間［ｓ，ｔ］（ｓ≦ｔ）の総数はN_y（N_y+1）／２個ある。各区間［ｓ，ｔ］は１つの整数ｐに一対一対応させる。

【００６９】例えば、次の関数ｈ（ｓ，ｔ）によって区間［ｓ，ｔ］を１つの整数ｐ＝ｈ（ｓ，ｔ）に対応させることができる。すなわち、

【数３８】

$$h(s,t)=S+\frac{t(t+1)}{2}$$

である。

【００７０】逆に、区間を表す整数ｐ（０≦ｐ≦N_y（N_y+1）／２）から区間［ｓ，ｔ］は、次のように求められる。

【数３９】

$$t=\left\lfloor\frac{-1+\sqrt{1+8p}}{2}\right\rfloor$$

【数４０】

$$S=P-\frac{t(t+1)}{2}$$

但し、このような計算はどの領域が最大のゲインを有するのかを決定する際には用いることはない。以後、［ｓ，ｔ］は１つの整数と同一視して取り扱う。また、上記数３８は一例にすぎず、他の関数を用いても問題ない。

【００７１】次に領域を記憶しておく配列を用意する。これは、$N_x\times N_y$（N_y+1）／２の整数型２次元配列であり、Ｗ，Ｕ，Ｄ，Ｎ－Ｔｙｐｅのそれぞれに対して１つ用意する。この要素$H^X(m,[s,t])$（$0\leqq m\leqq N_x-1$，$0\leqq[s,t]\leqq N_y(N_y+1)/2$，X∈｛W，U，D，N｝）と表すこととする。

【００７２】この要素$H^X(m,[s,t])$には、ゲインが$f_m^X(s,t)$の領域の第ｍ－１列の区間［ｘ，ｙ］と、第ｍ－１列の第ｍ－２列からの変化傾向Ｙを表す数値を記憶する。以下、$H^X(m,[s,t])$＝Ｙ：［ｘ，ｙ］と表す。例えば、この$H^X(m,[s,t])$を整数型３２ビットで表現し、Ｙの部分を上位２ビット、残りの下位ビットを［ｘ，ｙ］を表すのに用いる（図１８参照）。

【００７３】但し、第ｍ列が領域の左端列である場合に、第ｍ－１列にはつながらないことを表すために、この下位ビットには領域の左端を表す値を入れる。例えば、先ほどの数４０で区間を表現する例では、この下位ビットにN_y（N_y+1）／２以上の値を入れるか又は２９ビット目を領域の左端を表すフラグにすればよい。

【００７４】では、最終的に最大のゲインを有する直交凸領域を求める処理を図１９を用いて説明する。ステップ１６００で開始された処理は、最初にｍ＝０として、ｍを初期化する（ステップ１６１０）。次に、ｍ＝N_xであるか判断する（ステップ１６２０）。これは、ｍがN_xに達して、全ての列について以下の計算が終了したかを判断するものである。もし、全ての列ｍについて計算が終了していなければ、全ての［ｓ，ｔ］について、$H^W(m,[s,t])$と$f_m^W(s,t)$、$H^U(m,[s,t])$と$f_m^U(s,t)$、$H^D(m,[s,t])$と$f_m^D(s,t)$、$H^N(m,[s,t])$と$f_m^N(s,t)$を計算し、その結果を記憶する。この計算の順番は任意である。そして、各計算中それまでに計算されたゲインの最大値より大きい値が計算されたならば、その値及びそのｍ，［ｓ，ｔ］，Ｘを記憶しておく（ステップ１６３０）。

【００７５】ここで、$H^W(m,[s,t])$と$f_m^W(s,t)$の計算は、先に示した数３０の計算を実施すればよい。よって、$H^W(m,[s,t])$は、数３０の（１）式が最大であれば領域の左端を表す値、（２）式が最大であればＷ：［ｓ，ｔ］、（３）式が最大であれば$H^W(m,[s,t-1])$、（４）式が最大であれば$H^W(m,[s+1,t])$となる。以上のように、$H^W(m,[s,t])$のみを考えれば、前列である第ｍ－１列は、必ずＷ－Ｔｙｐｅであるから、第ｍ－１列の第ｍ－２列からの変化傾向は記憶する必要ない。

【００７６】また、$H^U(m,[s,t])$と$f_m^U(s,t)$の計算は、先に示した数３３の計算を実施すればよい。よって、$H^U(m,[s,t])$には、数３３の（１）式が最大であれば領域の左端を表す値、（２）式が最大であればＷ：［$\beta_{m-1}^W(s,t)$，ｔ］、（３）式が最大であればＵ：［$\beta_{m-1}^U(s,t)$，ｔ］、（４）式が最大であれば$H^U(m,[s,t-1])$が記憶される。

【００７７】$H^D(m,[s,t])$と$f_m^D(s,t)$の計算は、先に示した数３６の計算を実施すればよい。よって、$H^D(m,[s,t])$には、数３６の（１）式が最大であれば領域の左端を表す値、（２）式が最大であればＷ：［ｓ，$\tau_{m-1}^W(s,t)$］、（３）式が最大であればＤ：［ｓ，$\tau_{m-1}^D(s,t)$］、（４）式が最大であれば$H^D(m,[s+1,t])$が記憶される。

【００７８】最後に、$H^N(m,[s,t])$と$f_m^N(s,t)$の計算は、先に示した数３７の計算を実施すればよい。よって、$H^N(m,[s,t])$は、数３７の（１）式が最大であれば領域の左端を表す値、（２）式が最大であればＷ：［ｓ，ｔ］、（３）式が最大であればＵ：［ｓ，ｔ］、（４）式が最大であればＤ：［ｓ，ｔ］、（５）式が最大であればＮ：［ｓ，ｔ］、（６）式が最大であればＨ

$H^{N}(m,[s,t+1])$、（７）式が最大であれば$H^{N}(m,[s-1,t])$となる。

【００７９】ここまでで分かるように、すべての$f^{X}(s,t)$を記憶しておく必要はない。第m列の計算を実施している時には、その第m列と第m－１列の計算結果のみを用いる。よって、W，U，D，N－Ｔｙｐｅごとに２列分の記憶容量があればよい。但し、余裕があれば全て記憶しておいてもよい。

【００８０】図１９のステップ１６３０を終了すると、mを１インクリメントして（ステップ１６４０）、ステップ１６２０に戻る。そして、この処理を全ての列について実施する。もし、全ての列について実施されたならば、全ての列に関して最大のゲイン値を有していた領域に関するm，［ｓ，ｔ］，Xから、$H^{X}(m,[s,t])$を参照し、その値Ｙ：［ｘ，ｙ］を取り出す（ステップ１６５０）。ここまでの処理で、最右端列である第m列と、その列の区間［ｓ，ｔ］、第m－１列とその区間［ｘ，ｙ］が分かる。

【００８１】次に、第m－１列で領域は左端となる場合もあるので、［ｘ，ｙ］が左端を表す値であるか判断される（ステップ１６６０）。左端であれば、ここで処理は終了する（ステップ１６８０）。左端でなければ、YをXとして、［ｘ，ｙ］を［ｓ，ｔ］として、m－１をmとし（ステップ１６７０）、ステップ１６５０に戻る。このように、［ｘ，ｙ］が左端を表す値となるまで、この処理を繰り返せば、最大のゲイン値を有する直交凸領域の各列の区間を得ることができる。

【００８２】（３）出力ステップ

以上のように求まった直交凸領域Ｓは、前記平面のどの部分を占めているかは、先のステップによりわかっているので、その領域Ｓに属するデータを取り出すことになる。通常各データは、真偽をとる属性及び数値属性のみならず、他の属性も有しているから、例えばダイレクトメールを送るのであれば、住所氏名といった属性を取り出すようになる。ここまでくると、取り出すべきデータは特定されているから、通常のデータベースの検索に過ぎないので、これ以上詳しく述べない。当然、一旦直交凸領域をその外形がよくわかるようにして、ユーザに提示するようにしてもよい。

【００８３】以上のような各ステップを実施すれば、ある条件θに対する、データ間結合ルールの１つを求めることができる。しかし、この条件θをどのように設定するかということは、１つの問題である。通常、ある条件θ１つでは、問題の解決にならない場合が多い。以上の各ステップ、特に（２）領域切り出しステップをエンジンとして用い、どのように先に述べた４つの一般的なルール及び他のルール等を導き出すかを以下に示す。

【００８４】Ａ．ある区間に存在する直交凸領域を求める場合

まず、幾つかのθに対応するフォーカス・イメージＳを連続的に示し、動画を作成することにより、切り出される領域の大きさ及び形状をユーザの判断により決定させる場合を考える。

【００８５】この処理を図２０に示す。ステップ８００にて開始された処理は、まずθ１を入力することにより、上述したプロセスにてフォーカス・イメージＳ１を見つけ出す（ステップ８１０）。また、ユーザにθ２を入力させ、同様にフォーカス・イメージＳ２を見つけ出す（ステップ８２０）。このようにして２つのフォーカス・イメージが求まると、それぞれに含まれるデータ数Ｕ（Ｓ1），Ｕ（Ｓ2）及び真偽をとる属性が真であるデータの数Ｖ（Ｓ1），Ｖ（Ｓ2）とを用いて、その中間にある、新たな傾きθ３を計算する（ステップ８３０）。

【００８６】このように新たなθ３が求まれば、さらにこのθ３に対応するフォーカス・イメージＳ３を求めることができる（ステップ８４０）。ここで、計算されたＳ３が既に求まっていれば、区間（θ１，θ２）にはこれ以上のフォーカス・イメージは凸包上（図４）には存在しない。よって、処理が終了する（ステップ８８０）。しかし、発見済みでなければ、θ２の代わりにθ３を用いて、ステップ８３０以降を実行する（ステップ８６０）。すなわち、区間（θ１，θ３）の間にあるフォーカス・イメージを見つけ出す。この場合、次々に中間の値を計算していくようにすることも可能である。また、ある程度の個数フォーカス・イメージが求まったところで計算を取り止めることもできる。さらに、もう１つ残った区間（θ３，θ２）についてフォーカス・イメージを計算するために、θ３，θ２についてステップ８３０以降を実行する（ステップ８７０）。この場合も、この区間内に存在しているフォーカス・イメージを全て見つけ出すようにしてもよいし、所定の個数見つけ出したところで処理を終了してもよい。

【００８７】このようにして、１つ又は複数のフォーカス・イメージを見つけ出すことができた。このように求まった複数のフォーカス・イメージを連続してユーザに提示するようなことも可能である。

【００８８】Ｂ．コンフィデンス最大化ルールの場合（図２１及び図２２）

この場合には、ルールの定義より最小限度のサポートminsup（全体のデータ数に対する領域に包含されるデータ数の割合）を入力する（ステップ９１０）。ここで、Umin＝Usum×minsupを計算しておく。ここで図４を見てみると、最小限度サポートと記された縦の点線がこの値に対応する。まず、θ＝１でフォーカス・イメージＳ1を求める（ステップ９２０）。そして、このＳ1に含まれるデータ数Ｕ（Ｓ1）が、Ｕ（Ｓ1）＞Uminを満たすかどうか判断する（ステップ９３０）。もし成立するならば、Ｓ1を解として決定し（ステップ９５０）、処理を終了する（ステップ９９０）。成り立たない場合、フォーカス・イメージＳ2を平面全体を表すイメージとす

る。すなわち、U（S2)＝Usum，V（S2）＝Vsumと代入する（ステップ９４０）。そして、XXを介して図２２に移行する。

【００８９】図２２では、XXから始まり、新たな条件θを求め、このθに対するフォーカス・イメージSを計算する（ステップ１４００）。このθは

θ＝（V（S2）－V（S1））／（U（S2）－U（S1））

にて計算される。そして、S1＝S又はS2＝Sであるならば、（S1，S2）の間にはこれ以上フォーカス・イメージは存在しないので、コンフィデンスの高いS2が最良解として出力され、処理を終了する（ステップ１４１０）。また、U（S）≒Uminであるならば、Sを出力し、処理を終了する。

【００９０】ところが、U（S）＜Uminであると（ステップ１４２０）、まだ処理が必要なので、S1＝Sとして（ステップ１４４０）、ステップ１４００に戻る。同様に、U（S）＞Uminであるならば、S2＝Sとして（ステップ１４３０）、ステップ１４００に戻る。

【００９１】これを繰り返すことにより解が見つけられる。図４を参照すると、先に説明した最小限度のサポートの右側、濃く塗られた部分に解の存在する範囲がある。そして、この図４の場合には、凸包の内部の白丸の点が厳密解となるが、本発明ではハンド・プローブにて得られた近似解が出力される。見つけられた解は、ユーザに提示されるようにしてもよいし、そのフォーカス・イメージに属するデータの必要な属性を出力するようにしてもよい。

【００９２】C．サポート最大化ルールの場合（図２３，図２４）

この場合、ルールの定義より、最小限度のコンフィデンスminconf（直交凸領域に包含されるデータ数に対する真偽をとる属性が真である割合）を入力する（ステップ１１１０）。図４の場合、最小限度のコンフィデンスと示され、原点から引かれた点線がこれに該当する。まず、フォーカス・イメージS2を平面全体を表すイメージとする。すなわち、U（S2）＝Usum，V（S2）＝Vsumと代入する（ステップ１１２０）。そして、minconf≦V（S2）／U（S2）であるかを判断する（ステップ１１３０）。もしこの条件が成立するならば、S2を解として決定し（ステップ１１６０）、処理を終了する（ステップ１１９０）。条件が成立しないならば、θ＝１でフォーカス・イメージS1を求める（ステップ１１４０）。そして、minconf＞V（S1）／U（S1）が成り立つかどうか判断する（ステップ１１５０）。もし成り立つならば、解は存在せず、処理を終了する。成り立たないならば、Yを介して図２４へ移行する。

【００９３】図２４では、Yから処理が開始され、θ＝（V（S2）－V（S1））／（U（S2）－U（S1））としてフォーカス・イメージSを求める（ステップ１２００）。この求められたフォーカス・イメージSに対し、（1）minconf≒V（S）／U（S）が成立する場合には、このSを出力して処理を終了する（ステップ１２１０）。また、S1＝S若しくはS2＝Sである場合には、これ以上S1とS2の間には解は無いので、S1を最良解として出力し、処理を終了する（ステップ１２１０）。これに対し、minconf＜V（S）／U（S）である場合には（ステップ１２２０）、S1＝Sとしてステップ１２００に戻る（ステップ１２３０）。また、minconf＞V（S）／U（S）である場合には、S2＝Sとしてステップ１２００に戻る（ステップ１２４０）。

【００９４】以上のようにして、サポート最大化ルールが求められる。もう一度図４に戻ると、先に説明した最小限度のコンフィデンスとして示した点線より上の濃く塗られた範囲に解が存在する。そして、この例では凸包内の白丸の点が厳密解であるが、このように凸包内部の点は見つけ出すのに膨大な計算量を必要とするので、凸包上の点でサポートを最大にする近似解を出力するようにしている。先に述べたように、見出された近似解又は厳密解は、ユーザに提示してもよいし、フォーカス・イメージ内に含まれるデータの必要な属性値を出力するようにしてもよい。

【００９５】D．最適化エントロピ・ルールの場合

最適化エントロピ・ルールとは、領域の内部と外部との分割を考えた時、分割前の情報量と比較した分割後の情報量の増分を最大化するルールである。よって、切り出された領域と平面全体のエントロピのゲイン（以下の式）が最大となる領域を発見すればよい。

【数４１】

$$f(x,y) = -\frac{y}{a}\log\frac{y}{a} - \frac{x-y}{a}\log\frac{x-y}{a} - \frac{b-y}{a-x}\log\frac{b-y}{a-x} - \frac{a-b-x+y}{a-x}\log\frac{a-b-x+y}{a-x}$$

このxはU（S），yはV（S），aはUsum，bはVsumである。このような条件においても、解は凸包上に存在することが分かったので、上述のステップを用いることができる。よって、θを変化させ、数４１を最大化するフォーカス・イメージを求めればよい。

【００９６】E．最適化インタクラスバリアンス・ルールの場合

先に述べたように最適化インタクラスバリアンス・ルールとは、領域内外の分割を考えた時、内外の「標準化された真偽の割合の平均からのずれ」の二乗和を最大化するルールである。よって、切り出された領域と平面全体のインタクラスバリアンス（以下の式）が最大となる領域を発見すればよい。

【数４２】

33

$$f(x,y)=x\left(\frac{y}{x}-\frac{b}{a}\right)^2+(a+x)\left(\frac{b-a}{a-x}-\frac{b}{a}\right)^2$$

ｘ，ｙ，ａ，ｂは上述したものと同じである。このような条件においても、解は凸包上に存在することが分かったので、上述のステップを用いることができる。よって、θを変化させ、数４２を最大化するフォーカス・イメージを求めればよい。

【００９７】Ｆ．その他

以上述べたように、Ｕ（Ｓ）とＶ（Ｓ）上の凸包上の点に存在する又は存在すると近似できる場合には、上述したステップを用いれば高速にルールに該当する領域を導き出すことができる。

【００９８】Ｇ．二次的なルールの抽出

上述のプロセスを用いて１つのルールを見い出した後に、二次的なルールを見つけ出すことができる。すなわち、切り出した１のフォーカス・イメージに属するｖ(i,j)を除去し、ｖ(i,j)／ｕ(i,j)＝Ｖsum／Ｕsumとなるように、ｖ(i,j)を変更し、それから新たに領域切り出しステップを行うのである。

【００９９】以上、本発明における処理のプロセスを説明した。このような処理プロセスは、コンピュータ・プログラムによって実現し、実行するようにしてもよい。例えば、図２５のような通常のコンピュータ・システムにおいて実行できるようなプログラムにすることもできる。処理プログラムは、ＨＤＤ１０５０に格納され、実行時にはメインメモリ１０２０にロードされ、ＣＰＵ１０１０によって処理される。また、ＨＤＤ１０５０はデータベースをも含んでおり、処理プログラムはそのデータベースに対するアクセスを行う。最初の平面やフォーカス・イメージは、表示装置１０６０によってユーザに提示される。ユーザは、入力装置１０７０にてフォーカス・イメージの選択や、データ出力の命令を入力する。このような入力装置には、キーボードやマウス、ポインティング・デバイスやディジタイザを含む。さらに、出力結果を補助記憶装置であるＦＤＤ１０３０のフロッピー・ディスクに記憶したり、また新たなデータをＦＤＤ１０３０から入力することもできる。さらに、ＣＤ－ＲＯＭドライブ１０４０を用いて、データを入力することもできる。

【０１００】さらに、本発明の処理プロセスを実現したコンピュータ・プログラムは、フロッピー・ディスクやＣＤ－ＲＯＭといった記憶媒体に記憶して、持ち運ぶことができる。この場合、通常のデータベース検索プログラムのデータ取り出し部分や、表示装置１０６０に表示するだけの処理を行うプログラムは、すでにＨＤＤ１０５０に記憶されている場合もある。よって、それ以外の部分が、上記のような記憶媒体にて流通することは通常行われる事項である。また、図示されていない通信装置がバス１０８０に接続されており、遠隔地にあるデータ

34

ベースを用いて処理したり、処理結果を遠隔地に送信するようにしてもよい。

【０１０１】また、本発明の処理を実施する特別の装置を設けてもよい。例えば、図２６のような装置が考えられる。平面構成装置１３１０は、データベース１３００及び切出装置１３２０に接続されており、制御装置１３４０からの命令を受付ける。また、切出装置１３２０は、出力デバイス１３３０及び表示装置１３５０に接続されており、制御装置１３４０からの命令を受付ける。また、切出装置１１３０はデータベースにも接続を有している。制御装置１３４０は、入力デバイス１３６０に接続され、入力デバイス１３６０により指示された処理の種類により平面構成装置１３１０及び切出装置１３２０を制御する。

【０１０２】この装置の簡単な動作を説明する。平面構成装置１３１０は、先に説明した平面構成ステップを実行する部分である。このように平面構成装置１１１０は、データベースに記憶されたデータを用いて先に示した平面を構成し、切出装置１３２０に出力する。切出装置１３２０は、制御装置１３４０からの命令に従って、切り出しのためのパラメータであるθをセットする。セットされたθに従って切出装置１３２０は、先に述べた切出ステップを行い、フォーカス・イメージを切り出す。そして、表示装置１１４０に出力し、ユーザに命令されれば、切り出されたフォーカス・イメージ内に属するデータをデータベース１３００から取り出し、出力デバイス１３３０に引き渡す。出力デバイス１３３０は、適当な形式でユーザ所望のデータを出力する。また、ユーザは、例えば入力デバイス１３６０からコンフィデンス最大化ルールを解くように命じ、最小限度のサポートを入力する。すると、制御装置１３４０は先に示した処理Ｂを行うように、条件θを設定し、切出装置１３２０に出力する。そして、命じられたコンフィデンス最大化ルールに合致するような領域を解くべく、条件θを変化させる等の処理を行う。先に述べたサポート最大化ルール（処理Ｃ）や、最適化エントロピ・ルール（処理Ｄ）、最適化インタクラスバリアンス・ルール（処理Ｅ）、その他凸包上に位置する領域を切出す処理Ｆに適した条件θを切出装置１３２０に渡す処理を制御装置１３４０は行う。ユーザは入力デバイス１３６０から処理の種類や、先に述べたような条件（θのみならず、minconf，minsupも）を入力する。また、制御装置１３４０は、上述の処理Ｇを行うために平面構成装置１３１０に、切り出したフォーカス・イメージのＶ(i,j)を除去する等の処理を命じる。

【０１０３】以上、本発明を特別の装置にする一例を示したが、本発明はこれに限定されるものではない。例えば、切出装置１３２０の出力は、出力制御装置を介して出力デバイス１１６０及び表示装置１１４０に出力されるようにしてもよいし、この場合出力制御装置からデー

タベースを参照してデータを取り出すようにしてもよい。

【０１０４】以上は、通常データが有するＫ個の数値属性のうち２項を選択し、それらの数値属性間の壮観を見つける処理であったが、数２３を目的関数とし、ｎ次元空間の領域を切り出すことができれば、ｎ次元の探索に拡張することができる。

【０１０５】

【効果】２項以上の数値属性と真偽をとる属性を有するデータ間の結合ルールを見い出すための一手法を提供することができた。

【０１０６】また、データ間の結合ルールを人間がより把握しやすい形で提示することもできた。そして、多くの結合ルールを可視化することにより、使用する人間の選択の幅を増大させ、より重要な結合ルールを見いだすこと可能とすることもできた。

【０１０７】さらに、（１）サポート最大化ルールや、（２）コンフィデンス最大化ルール、（３）最適化エントロピ・ルール、（４）最適化インタクラスバリアンス・ルールを満たすような範囲（領域）を導出可能とすることもできた。

【０１０８】また、上記のようなデータ間の結合ルールを高速に実行できるような手法を提供することもできた。

【０１０９】例えば、ある割合以上で、例えばアウトドアスポーツに興味を示す（真偽をとる属性に相当する）、できるだけまとまった領域に入る顧客を知ることができるので、その条件に合致する多くの顧客に知ってもらいたいダイレクトメールの宛て先を知るのに用いることができる。（サポート最大化ルール）

【０１１０】一定数以上の顧客を含む、例えば定期預金残高２００万円以上の顧客割合が最も高いところを知ることができるので、顧客を絞りこみつつ、有効な宣伝活動等を行うことができる。（コンフィデンス最大化ルール）

【図面の簡単な説明】

【図１】平面構成ステップのフローを示す図である。

【図２】領域切り出しステップのための前準備のフローを示す図である。

【図３】直交凸領域を説明するための図である。

【図４】Ｕ（Ｓ），Ｖ（Ｓ）平面の説明をするための図である。

【図５】領域切り出しステップにおける表記を説明するための図である。

【図６】第ｍ－１列から第ｍ列への変化傾向を説明するための図である。

【図７】直交凸領域の各列の状態遷移を表す図である。

【図８】Ｗ－Ｔｙｐｅにおける第ｍ列と第ｍ－１列の関係を説明するため図であって、（ａ）は数３０の（２）式、（ｂ）は（３）式、（ｃ）は（４）式を説明するための図である。

【図９】$f_m^W(s,t)$の計算順番を示した図である。

【図１０】Ｕ－Ｔｙｐｅの計算に用いる前処理のアルゴリズムを説明するための図である。

【図１１】Ｕ－Ｔｙｐｅにおける第ｍ列と第ｍ－１列の関係を説明するための図であって、（ａ）は数３３の（２）式、（ｂ）は（３）式、（ｃ）は（４）式を説明するための図である。

【図１２】$f_m^U(s,t)$の計算順番を示した図である。

【図１３】Ｄ－Ｔｙｐｅの計算に用いる前処理のアルゴリズムを説明するための図である。

【図１４】Ｄ－Ｔｙｐｅにおける第ｍ列と第ｍ－１列の関係を説明するための図であって、（ａ）は数３６の（２）式、（ｂ）は（３）式、（ｃ）は（４）式を説明するための図である。

【図１５】$f_m^D(s,t)$の計算順番を示した図である。

【図１６】Ｎ－Ｔｙｐｅにおける第ｍ列と第ｍ－１列の関係を説明するための図であって、（ａ）は数３７の（１）式、（ｂ）は（３）式、（ｃ）は（４）式、（ｄ）は（５）式、（ｅ）は（６）式、（ｆ）は（７）式を説明するための図である。

【図１７】$f_m^N(s,t)$の計算順番を示した図である。

【図１８】$H^X(m,[s,t])$のデータ構造を示すための図である。

【図１９】領域切り出しステップの処理フローを表す図である。

【図２０】複数のフォーカス・イメージを見つけ出す処理のフローを示す図である。

【図２１】コンフィデンス最大化ルールを導出するための処理の一部を示すための図である。

【図２２】コンフィデンス最大化ルールを導出するための処理の一部を示すための図である。

【図２３】サポート最大化ルールの導出するための処理の一部を示すための図である。

【図２４】サポート最大化ルールの導出するための処理の一部を示すための図である。

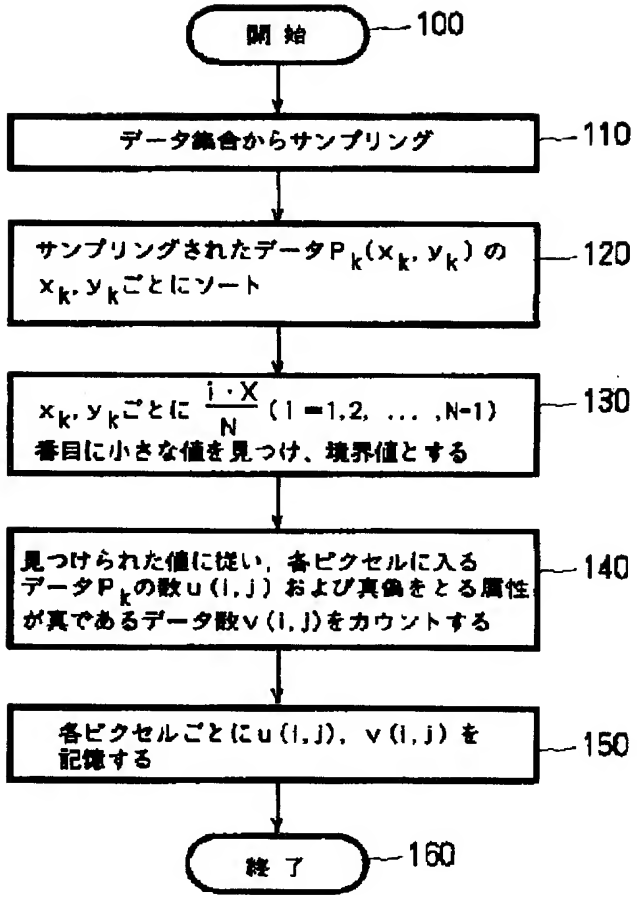
【図２５】通常のコンピュータ・システムで本発明を実施した場合の装置構成の一例を示す図である。

【図２６】本発明を専用の装置で実施した場合のブロック図である。

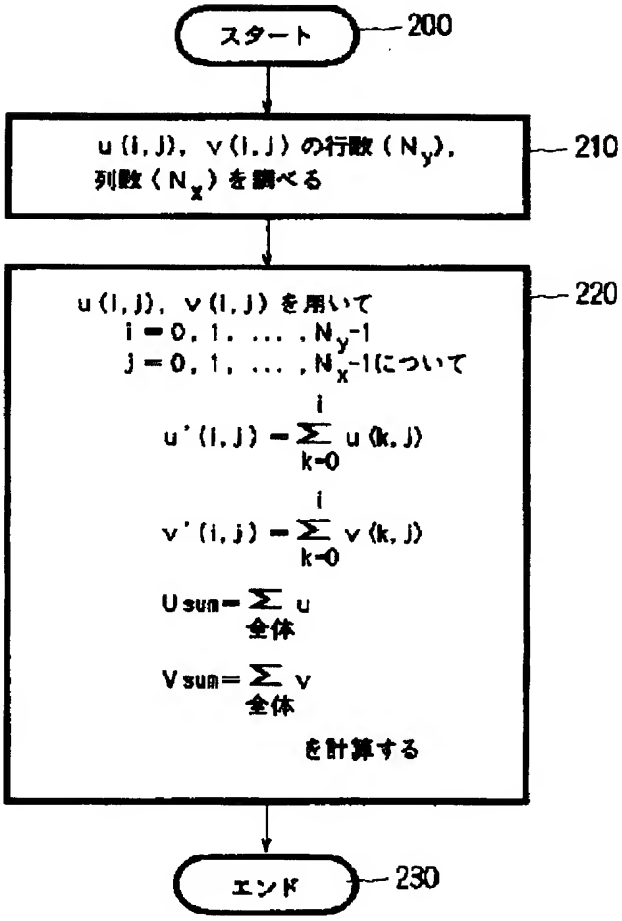
【符号の説明】

１０１０	ＣＰＵ	１０２０	メインメモリ
１０３０	ＦＤＤ	１０４０	ＣＤ－ＲＯＭドライブ
１０５０	ＨＤＤ	１０６０	表示装置
１０７０	入力デバイス		
１３１０	平面構成装置		
１３００	データベース	１３２０	切出装置
１３５０	表示装置	１１３０	入力デバイス
１３３０	出力デバイス	１３４０	制御デバイス

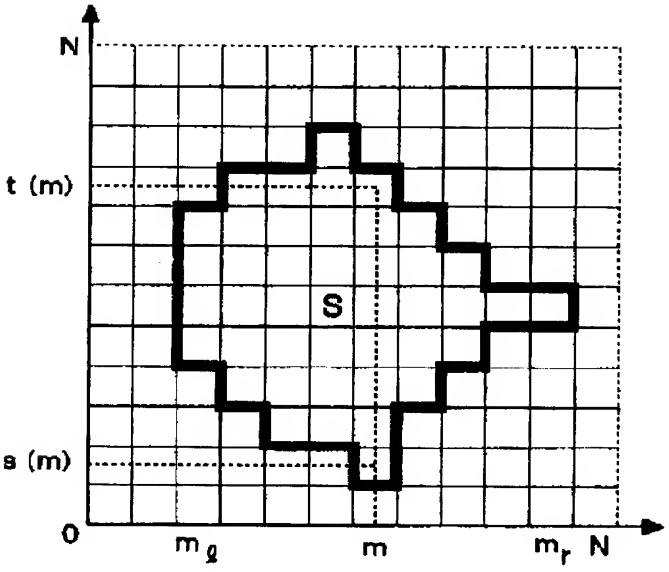
【図１】


開 始
100
データ集合からサンプリング
110
サンプリングされたデータ$P_k(x_k, y_k)$のx_k, y_kごとにソート
120
x_k, y_kごとに $\frac{i \cdot X}{N}$ (i=1,2, ... ,N-1) 番目に小さな値を見つけ、境界値とする
130
見つけられた値に従い、各ピクセルに入るデータP_kの数u(i,j)および真偽をとる属性が真であるデータ数v(i,j)をカウントする
140
各ピクセルごとにu(i,j), v(i,j)を記憶する
150
終 了
160

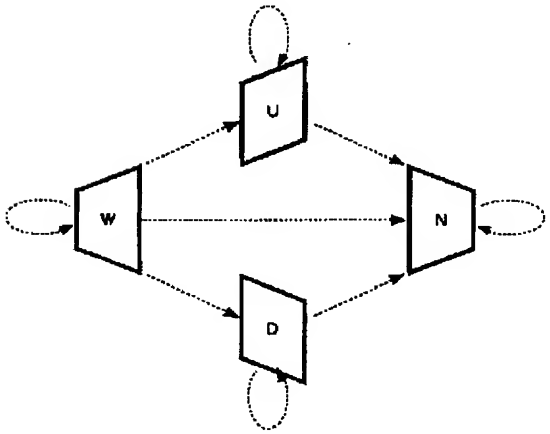
【図２】


スタート
200
u(i,j), v(i,j)の行数(N_y), 列数(N_x)を調べる
210
u(i,j), v(i,j)を用いて
i = 0, 1, ... , N_y-1
j = 0, 1, ... , N_x-1について
$u'(i,j) = \sum_{k=0}^{i} u(k,j)$
$v'(i,j) = \sum_{k=0}^{i} v(k,j)$
$U_{sum} = \sum_{全体} u$
$V_{sum} = \sum_{全体} v$
を計算する
220
エンド
230

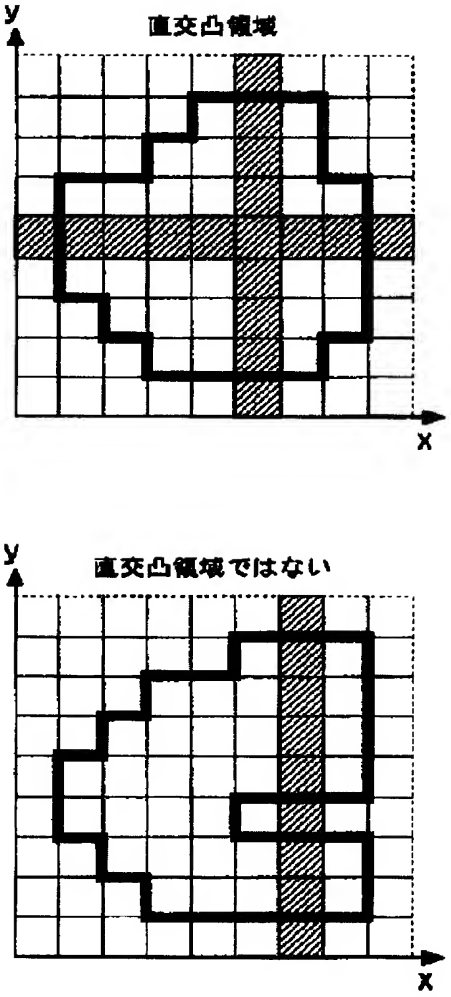
【図５】


N
t (m)
s (m)
0
S
m_l
m
m_r
N

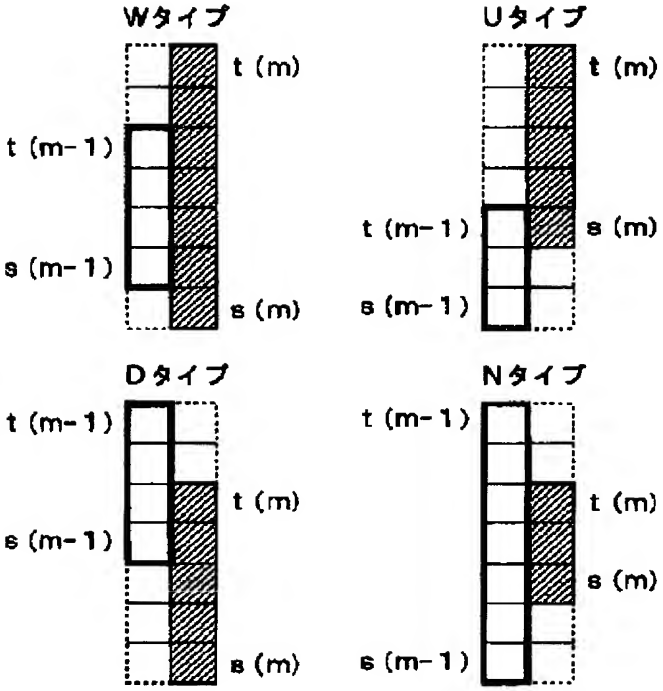
【図７】


U
W
N
D

【図３】



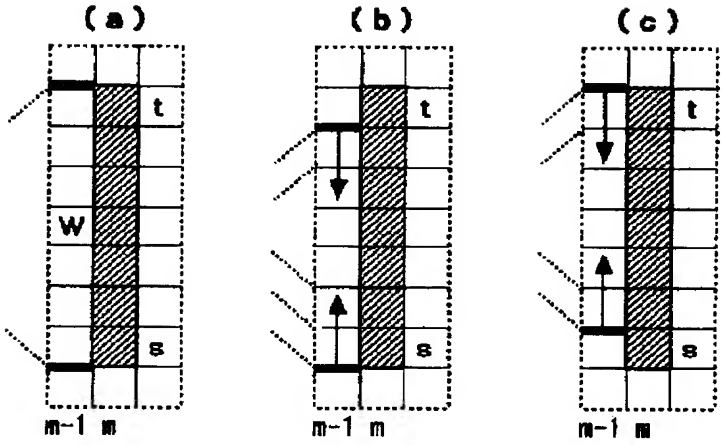
【図６】



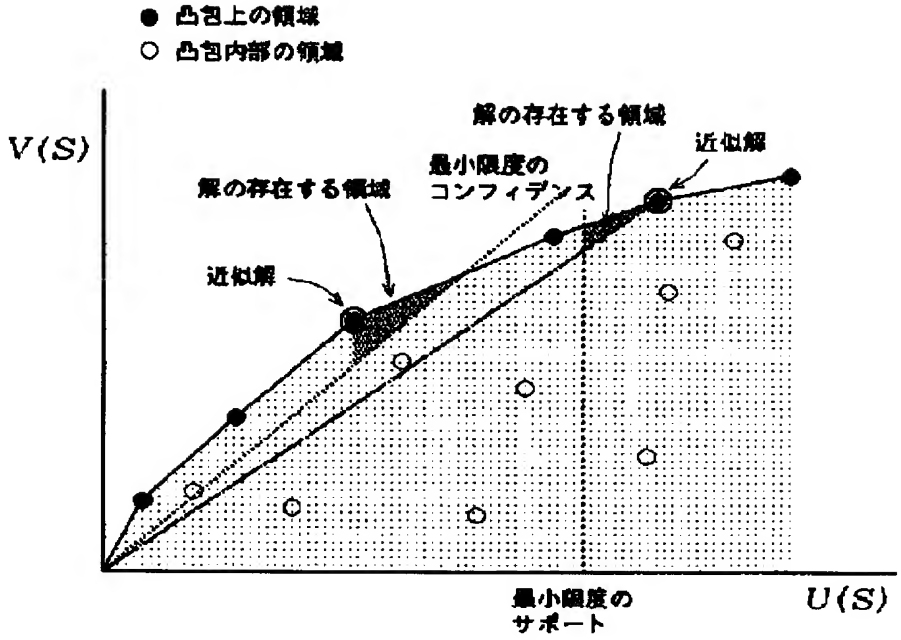
【図１８】

31, 30	29, ... , 0
変化傾向 Y	区間 [u, v]

【図８】



【図４】



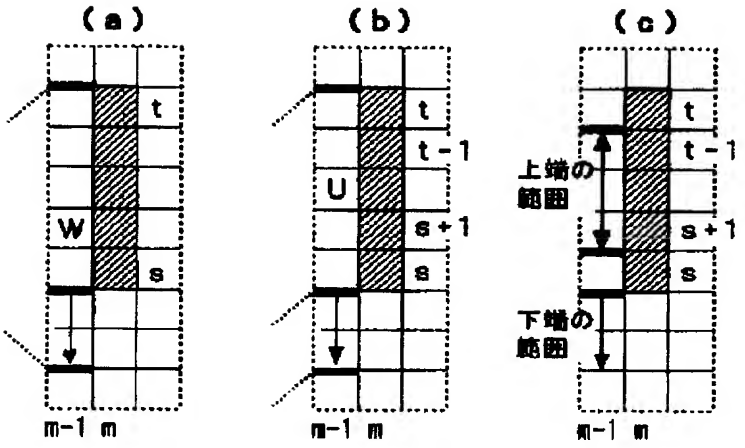
【図９】

```
for k=0 to N_y-1 do
        for s=0 to N_y-1-k do
                t=s+k
                f_m^W(s,t) を数３０を用いて計算
        end
end
```

【図１０】

```
for t:=0 to N_y-1 do
        β^W=0
        β^U=0
        for s:=0 to t do
                if f_{m-1}^W(s,t) > f_{m-1}^W(β^W,t) then β^W=s
                if f_{m-1}^U(s,t) > f_{m-1}^U(β^U,t) then β^U=s
                β_{m-1}^W(s,t)=β^W
                β_{m-1}^U(s,t)=β^U
        end
end
```

【図１１】



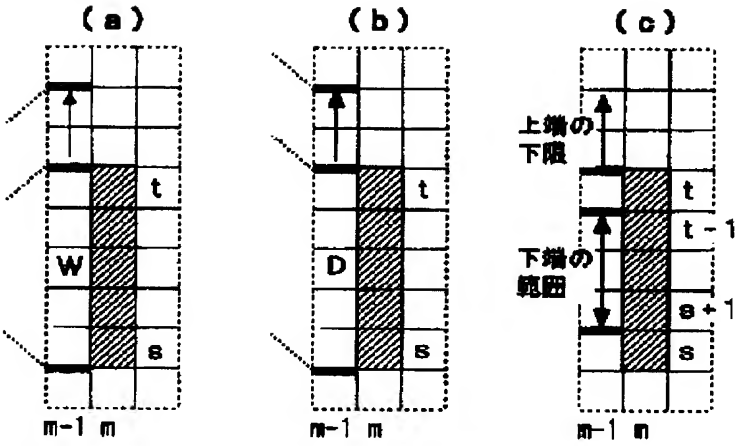
【図１２】

```
for s=0 to Ny-1 do
        for t=s to Ny-1 do
                f^U_m(s,t) を数３３を用いて計算
        end
end
```

【図１３】

```
for s:=Ny-1 to 0 do
        τ^W=Ny-1
        τ^D=Ny-1
        for t:=Ny-1 to s do
                if f^W_m-1(s,t) > f^W_m-1(s, τ^W) then τ^W=t
                if f^D_m-1(s,t) > f^D_m-1(s, τ^D) then τ^D=t
                τ^W_m-1(s,t)=τ^W
                τ^D_m-1(s,t)=τ^D
        end
end
```

【図１４】



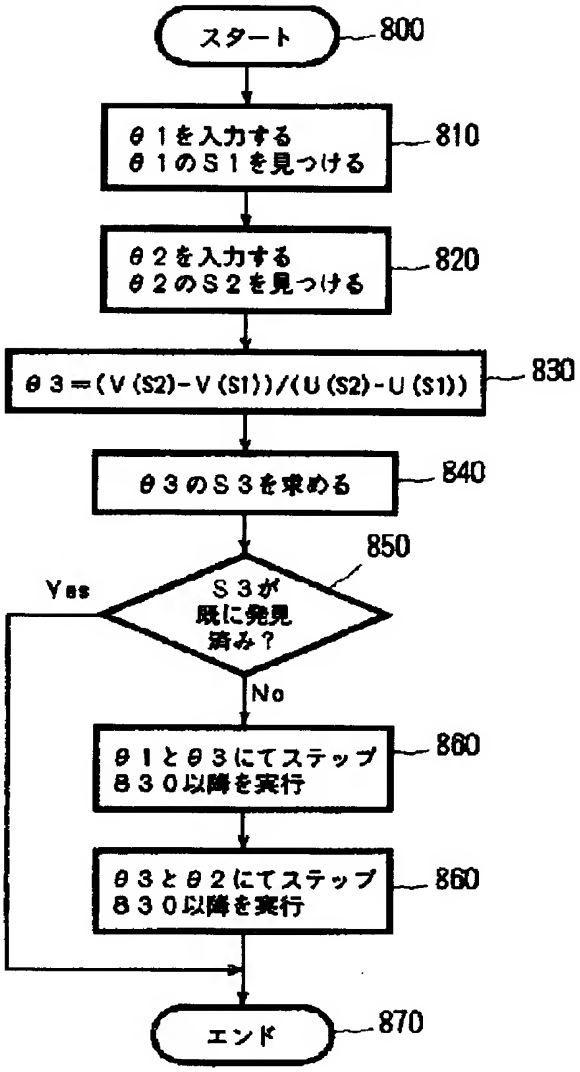
【図１５】

```
for t=N_y-1 to 0 do
        for s=t to 0 do
                f_m^D(s, t)　を数３６を用いて計算
        end
end
```

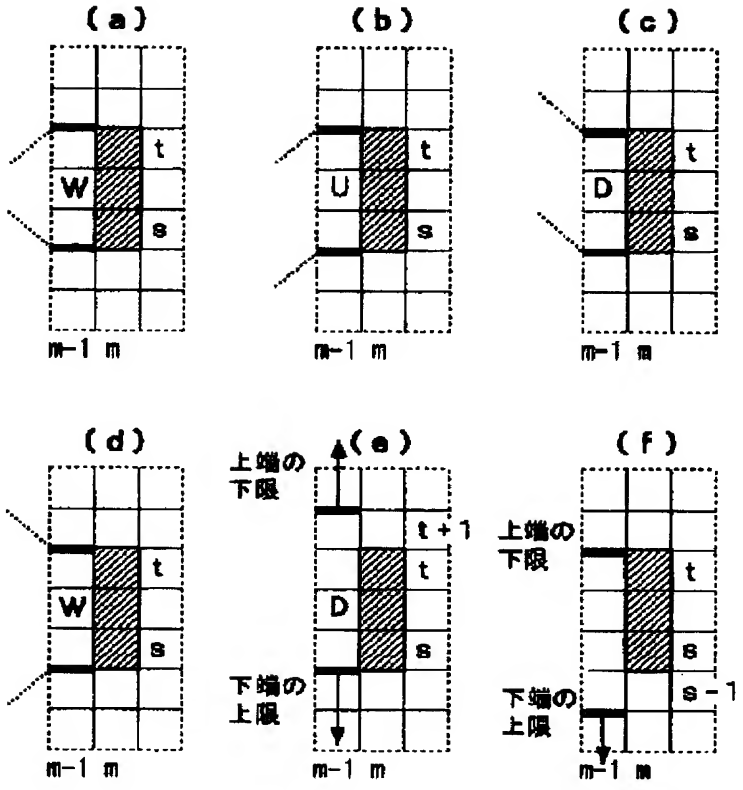
【図１７】

```
for k=N_y-1 to 0 do
        for t=N_y-1 to k do
                s=t-k
                f_m^N(s, t)　を数３７を用いて計算
        end
end
```

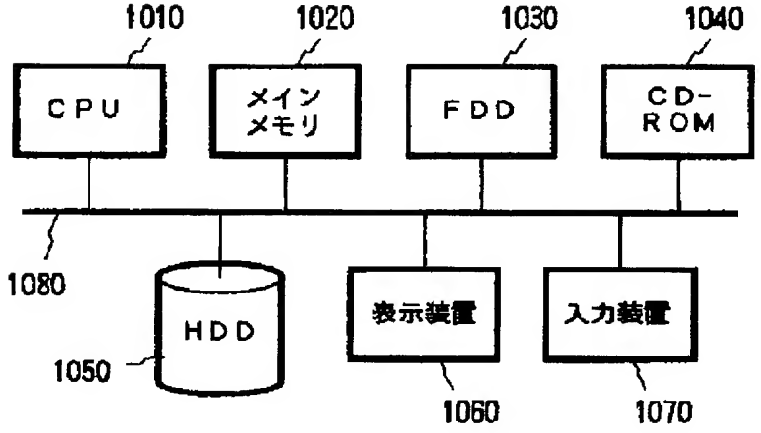
【図２０】



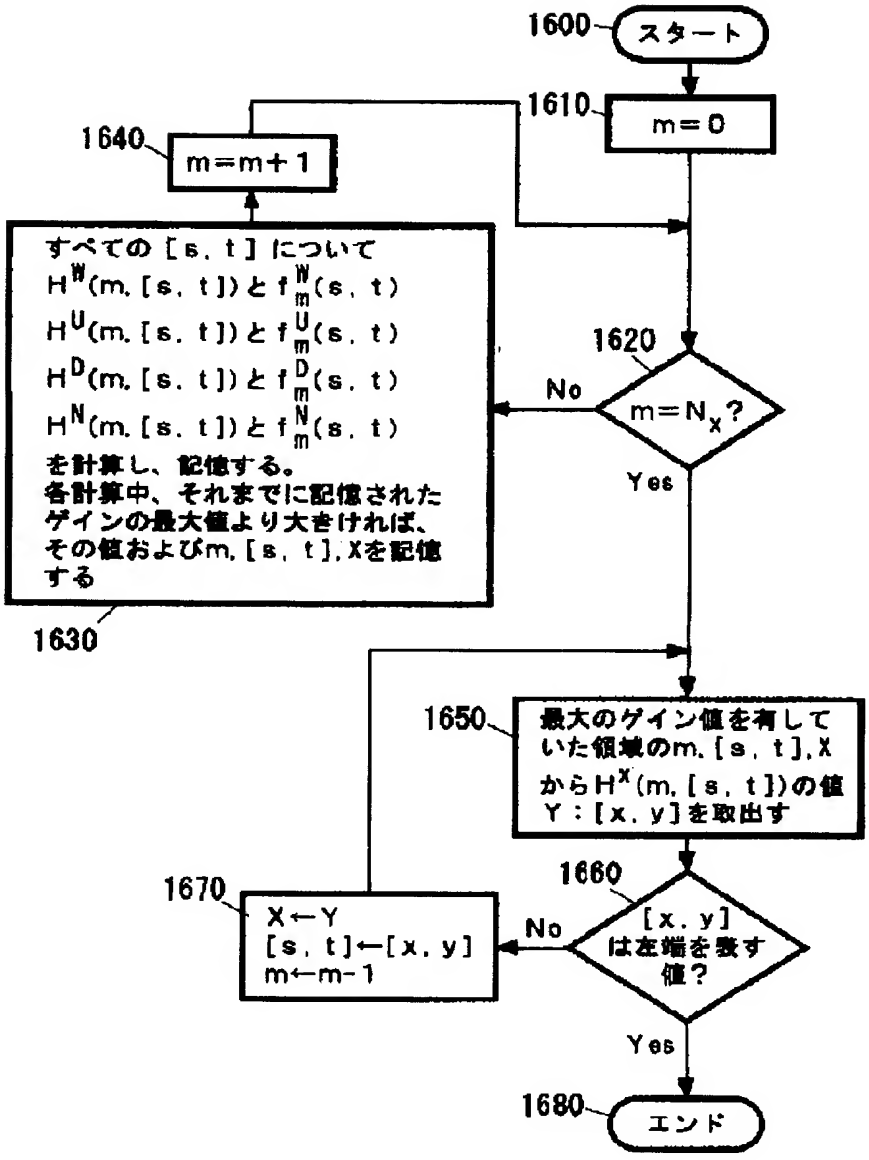
【図１６】


(a)
(b)
(c)
(d)
(e)
(f)
t
s
W
U
D
m-1 m
上端の下限
下端の上限
t+1
s-1

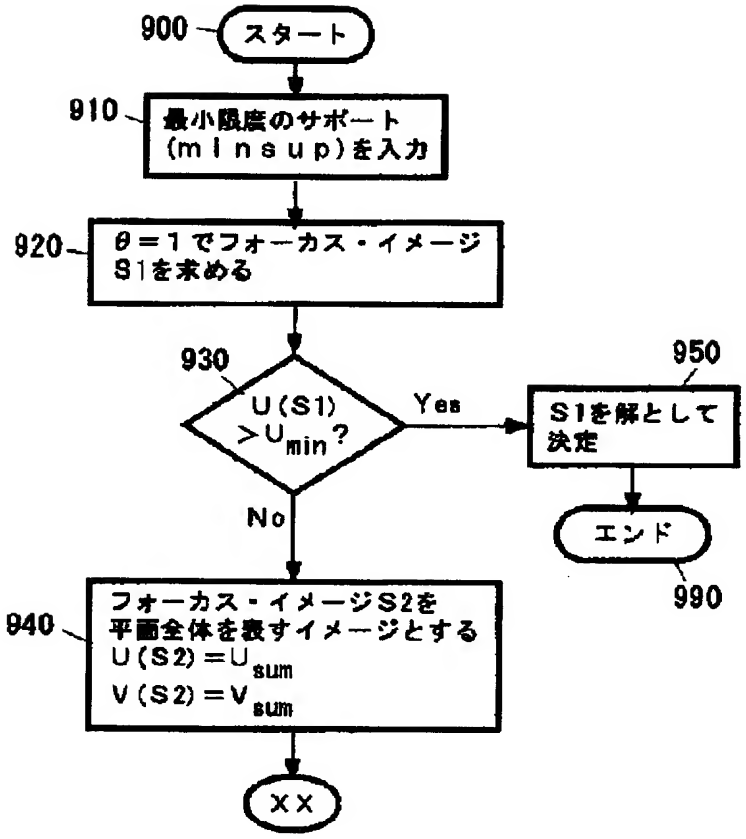
【図２５】


1010
1020
1030
1040
CPU
メイン メモリ
FDD
CD-ROM
1080
HDD
1050
表示装置
1060
入力装置
1070

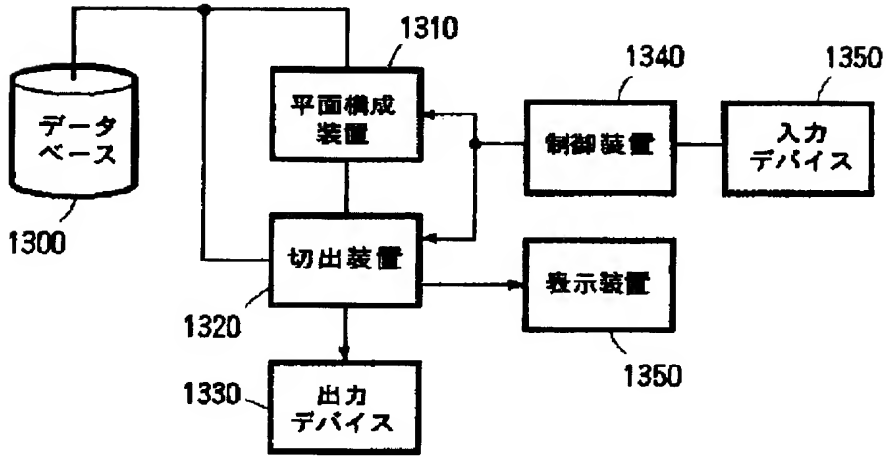
【図１９】


1600
スタート
1610
m＝0
1640
m＝m＋１
すべての［s，t］について
$H^W(m,[s,t])$と$f^W_{MEMBER}(s,t)$
$H^U(m,[s,t])$と$f^U_{MEMBER}(s,t)$
$H^D(m,[s,t])$と$f^D_{MEMBER}(s,t)$
$H^N(m,[s,t])$と$f^N_{MEMBER}(s,t)$
を計算し、記憶する。
各計算中、それまでに記憶されたゲインの最大値より大きければ、その値およびm，［s，t］，Xを記憶する
1630
1620
No
$m=N_X$?
Yes
1650
最大のゲイン値を有していた領域のm，［s，t］，Xから$H^X(m,[s,t])$の値Y：［x，y］を取出す
1660
［x，y］は左端を表す値？
No
1670
X←Y
［s，t］←［x，y］
m←m－１
Yes
1680
エンド

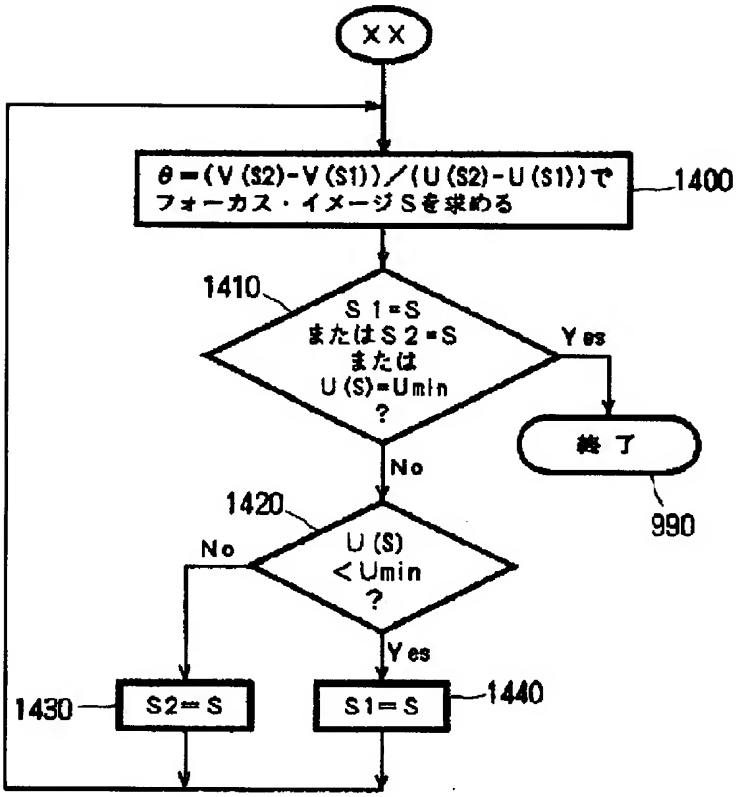
【図２１】


900
スタート
910
最小限度のサポート
(minsup)を入力
920
θ＝１でフォーカス・イメージ
S1を求める
930
U(S1)
>Umin?
Yes
950
S1を解として
決定
エンド
990
No
940
フォーカス・イメージS2を
平面全体を表すイメージとする
U(S2)＝Usum
V(S2)＝Vsum
XX

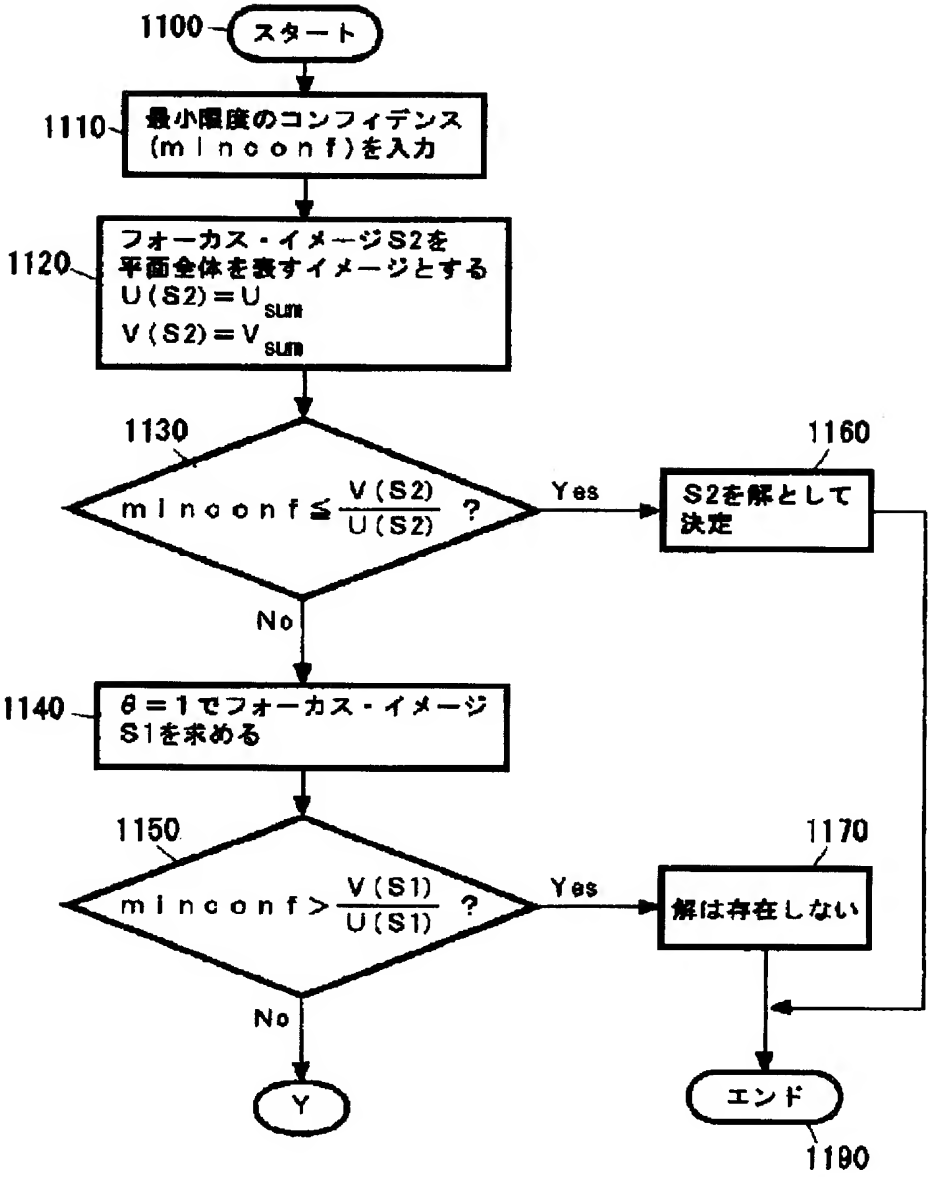
【図２６】


1310
データ
ベース
1300
平面構成
装置
1340
制御装置
1350
入力
デバイス
切出装置
1320
表示装置
1350
1330
出力
デバイス

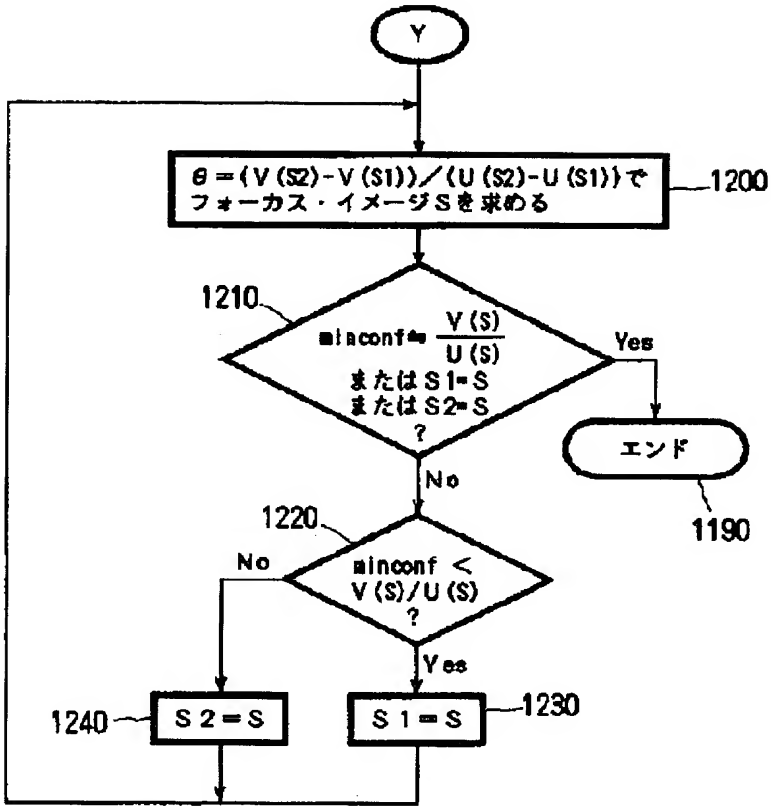
【図２２】


XX
θ＝(V(S2)-V(S1))／(U(S2)-U(S1))で
フォーカス・イメージSを求める
1400
1410
S1=S
またはS2=S
または
U(S)=Umin
?
Yes
終了
990
No
1420
No
U(S)
<Umin
?
Yes
1430
S2=S
S1=S
1440

【図２３】


1100
スタート
1110
最小確度のコンフィデンス
(minconf)を入力
1120
フォーカス・イメージS2を
平面全体を表すイメージとする
U(S2)＝U sum
V(S2)＝V sum
1130
minconf≦V(S2)/U(S2) ?
Yes
1160
S2を解として
決定
No
1140
θ＝1でフォーカス・イメージ
S1を求める
1150
minconf＞V(S1)/U(S1) ?
Yes
1170
解は存在しない
No
Y
エンド
1190

【図２４】



【手続補正書】
【提出日】平成９年１０月７日
【手続補正１】
【補正対象書類名】明細書
【補正対象項目名】００１８
【補正方法】変更
【補正内容】
【００１８】さらに、切り出された直交凸領域Ｓ内の各ピクセルのｖ(i,j)／ｕ(i,j)が、平面全体のデータ数に対する平面全体の真偽をとる属性が真であるデータ数の割合に等しくなるようｖ(i,j)を変更し、当該変更されたｖ(i,j)を用いて、入力された条件θ_4に従い、
【数２０】

$$\sum_{(i,j)\in S_4} g(i,j)=\sum_{(i,j)\in S_4}(v(i,j)-\theta_4 u(i,j))$$

を最大にするようなピクセルの第４の領域S_4を切り出すようにすることも考えられる。このようにすると、二次的な相関ルールを導き出すことができる。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】００１９
【補正方法】変更
【補正内容】
【００１９】また、先の平面構成ステップは、複数のデータから、Ｘ個のデータをランダムサンプリングし、サンプリングされたデータを各数値属性についてソートし、Ｘ・ｉ／Ｎ（ｉ＝１，２，・・Ｎ）番目に該当する数値及びＸ・ｎ／Ｍ（ｎ＝１，２，・・Ｍ）番目に該当する数値を記憶し、記憶された数値を基準にして、複数のデータをＮ×Ｍ個のピクセルに入れるようにすることも考えられる。このようにすると、各行各列にデータを高速にまたほぼ均等に割り振ることができる。
【手続補正３】
【補正対象書類名】明細書
【補正対象項目名】００７１
【補正方法】変更
【補正内容】
【００７１】次に領域を記憶しておく配列を用意する。これは、$N_x \times N_y$（N_y+1）／２の整数型２次元配列であり、Ｗ，Ｕ，Ｄ，Ｎ－Ｔｙｐｅのそれぞれに対して１つ用意する。この要素を$H^X(m,[s,t])$（$0 \leqq m \leqq N_x$

－１，０≦[s,t]≦N_y（N_y＋１）／２，X∈｛W，U，D，N｝）と表すこととする。
【手続補正４】
【補正対象書類名】明細書
【補正対象項目名】０１０４
【補正方法】変更
＊【補正内容】
【０１０４】以上は、通常データが有するＫ個の数値属性のうち２項を選択し、それらの数値属性間の相関を見つける処理であったが、数２３を目的関数とし、ｎ次元空間の領域を切り出すことができれば、ｎ次元の探索に拡張することができる。

フロントページの続き

(72)発明者　福田　剛志
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内
(72)発明者　徳山　豪
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内
(72)発明者　森下　真一
神奈川県大和市下鶴間1623番地14　日本アイ・ビー・エム株式会社　東京基礎研究所内